Facultat d'Informàtica de Barcelona

Grau en Enginyeria Informàtica

Especialitat d'Enginyeria del Software

Projecte de Final de Grau

Caracterització i detecció de textos generats artificialment

Casas Muñoz, Antoni

Dirigit per

Mario Martín Muñoz

26 de juny de 2020

Índex

1 Resum **7**

2 Introducció i contextualització **8**

2.1 Conceptes fonamentals 8

2.1.1 Conceptes lingüístics 9

2.2 Descripció del problema 11

2.3 Actors Implicats 15

3 Abast del projecte **16**

3.1 Solucions Existents 16

3.2 Justificació de la solució escollida 17

3.3 Objectius 17

3.4 Riscos i obstacles 18

3.5 Metodologia i rigor 19

3.5.1 Eines 19

4 Obtenció dels textos **20**

5 Anàlisi dels textos **23**

5.1 Anàlisi respecte a lleis cognitives 23

5.1.1 Anàlisi de la proporció de categories gramaticals 23

5.1.2 Hillberg's law 25

5.1.2.1 Text pur 26

5.1.2.2 Categories gramaticals 28

5.1.2.3 Caràcters 30

5.1.3 Zipf 33

5.1.4 Probabilitat mida paraula 34

5.2 Anàlisi de dades semàntiques 37

5.2.1 Distancies de dependències sintàctiques 37

5.2.2 Polisèmia 39

5.2.3 Coreferències 41

6 Precisió de classificació de textos 44

7 Conclusions del anàlisis 45

8 Disseny Software 46

8.1 Arquitectura 46

8.2 Documentació 46

8.3 Implementació 47

8.3.1 Llei de Hillberg 47

8.3.2 Llei de Zipf 48

8.3.3 Coreferències 48

8.3.4 Mida de paraula 48

8.3.5 Polisèmia 49

8.3.6 Distancies sintàctiques 49

9 Planificació Temporal 50

9.1 Definició de les tasques 50

9.1.1 Gestió del projecte 50

9.1.2 Anàlisi 52

9.1.3 Implementació d'API d'extracció d'informació 55

9.2 Diagrama de Gantt 56

9.3 Gestió de risc 56

10 Gestió econòmica 58

10.1 Costos per activitat 58

10.2 Costos genèrics 59

10.2.1 Amortitzacions 59

10.2.2 Espai de treball 60

10.2.3 Consum elèctric 60

10.2.4 Costos genèrics totals 60

10.3 Costos de contingència 61
10.4 Costos per imprevistos 61
10.5 Costos totals 61
10.6 Gestió del projecte 61

11 Sostenibilitat i compromís social 63
11.1 Estudi del impacte ambiental 63
11.2 Estudi de l'impacte econòmic 63
11.3 Estudi del impacte social 64

12 Conclusions 65
12.1 Treball Futur 65

13 Bibliografia 67

Índex de figures

1 Exemple dependències sintàctiques . . . 10
2 Text Original . . . 11
3 Text generat per GPT2 . . . 12
4 Pàgina de fake news completament artificial . . . 13
5 Creixement exponencial dels models transformer . . . 14
6 Precisió al diferenciar noticies escrites per humans i GPT3 . . . 15
7 Diferents gràfics representant diferents lleis de la lingüística quantitativa . . . 17
8 Text exemple de webtext . . . 21
9 Text exemple de GPT2 XL-1542 . . . 22
10 Proporció categories gramaticals de text humà . . . 24
11 Proporció categories gramaticals de GPT2 . . . 25
12 Diferencia de proporcions de categories gramaticals . . . 25
13 Entropia condicional del text pur humà . . . 26
14 Entropia condicional del text pur GPT2 . . . 27
15 Diferencia entre distribucions de entropia condicional sobre text pur entre text humà i GPT2 27
16 Divergència Jensen-Shannon entre distribucions de Hillberg sobre text pur en text humà i GPT2 . . . 28
17 Entropia condicional de les categories gramaticals del text humà . . . 29
18 Entropia condicional de les categories gramaticals del text GPT2 . . . 29
19 Diferencia entre distribucions de entropia condicional sobre categories gramaticals entre text humà i GPT2 . . . 30
20 Divergència Jensen-Shannon entre distribucions de Hillberg sobre categories gramaticals en text humà i GPT2 . . . 30
21 Entropia condicional de els caràcters del text humà . . . 31
22 Entropia condicional de els caràcters del text humà del text GPT2 . . . 32
23 Diferencia entre distribucions de entropia condicional sobre caràcters entre text humà i GPT2 32
24 Divergència Jensen-Shannon entre distribucions de Hillberg sobre caràcters en text humà i GPT2 . . . 33

25 Freqüència paraules en text humà 34
26 Freqüència paraules en text GPT2 34
27 Distribució mida paraula en text humà 36
28 Distribució mida paraula en text GPT2 36
29 Diferencia distribució mida paraules entre text humà i GPT2 37
30 Freqüència de distancies sintàctiques del text humà 38
31 Freqüència de distancies sintàctiques del text GPT2 38
32 Diferencia entre distribucions de distàncies sintàctiques entre text humà i GPT2 39
33 Distribució de nombre de synsets del text humà 40
34 Distribució de nombre de synsets del text artificial 40
35 Diferencia de distribució de nombre de synsets entre text humà i GPT2 41
36 Probabilitat de clusters de coreferències del text humà 42
37 Probabilitat de clusters de coreferències del text GPT2 42
38 Diferencia de distribucions de coreferències entre el text humà i GPT2 43
39 Diagrama de Gantt 56

Índex de taules

1 Precisió al classificar text entre humà o GPT2 44

2 Retribució horària per rol 58

3 Costos del personal basats en les tasques de la planificació temporal 59

4 Cost del hardware 60

5 Cost del consum elèctric 60

6 Costos genèrics del projecte 60

7 Cost dels imprevistos 61

8 Costos totals del projecte 61

1 Resum

Millores recents en el camp dels models de llenguatge natural han portat a la creació de nous models generadors de llenguatge, aquests nous models són de gran qualitat, i en certes ocasions, diferenciar-los d'allò que un humà escriuria o faria és extremadament complex [26]. A la vegada, usos il·legitims d'aquesta nova tecnologia estan creixent, per tant és d'interès la comprensió d'aquests models per a la seva millora, i per a la detecció d'usos il·legitims d'aquests.

Aquest treball examina diferents lleis i distribucions sobre el llenguatge natural, i examina quines diferències existeixen entre el text generat pel model màquina GPT2, el state of the art actual, i text escrit per humans. Específicament analitza la distribució de categories gramaticals, entropia condicional sobre el text, entropia condicional sobre les seves categories gramaticals, i entropia condicional sobre els caràcters del text, la distribució de zipf, la distribució de les mides de grups de correferències, la distribució de mides de paraula i la distribució de la polisèmia de cada paraula.

També s'ha desenvolupat una API REST documentada per Swagger 2.0 per a facilitar l'extracció d'aquesta informació i fer futurs anàlisis d'aquest estil més fàcils, i permetre la integració d'aquesta informació a processos d'extracció d'informació per l'avaluació de models de llenguatge natural creats amb aprenentatge màquina.

2 Introducció i contextualització

Recentment, la millora de la tecnologia i tècniques existents al camp de "Natural Language Processing", ha portat a una revolució sobre antics reptes abans considerats d'una complexitat molt elevada, com per exemple, traducció automàtica de textos.

Aquesta revolució ha provocat una ràpida expansió de tot el que relaciona al llenguatge natural, especialment en tasques de processament automàtic de llenguatge, on ara solucions que són més ràpides, simples i precises reemplacen ràpidament antics sistemes i models. D'especial interès aquí, és la generació automàtica de textos, que, debut al creixement de les fake news ha portat un escrutini especial gràcies al seu potencial de poder crear quantitats industrials de fake news automàticament.

2.1 Conceptes fonamentals

Els models de llenguatge son una distribució de probabilitats sobre seqüencies de paraules, existeixen diferents tipus d'aquests models, però els més importants son els models de llenguatge de xarxes neuronals. En aquest cas, el model intenta predir una distribució donat un context. D'aquests tipus de models de llenguatge, els més importants ara mateix son els transformers, que s'han transformat en models de llenguatge general, es a dir, un model pot fer diferents tasques, com traducció, predicció de la següent paraula, respondre preguntes, etc...

Antigament, la generació automàtica de textos, seguia sistemes basats en regles, on les millors paraules eren escollides per completar arbres sintàctics preexistents [21]. Aquestes solucions podien generar texts que eren suficientment expressius com per a comunicar conceptes, però aquests eren fàcils d'identificar com a automàtics, ja que al ser creat basat en regles, era simple trobar aquestes per identificar els textos.

Amb el temps, solucions que fusionaven mètodes basats en regles amb xarxes neuronals van aparèixer, però aquestes encara sofrien dels mateixos problemes que les solucions anteriors tenien, el text era molt rígid.

L'any 2017, però, Google va publicar [31], on un nou mètode per a la solució del problema de seqüència a seqüència, on una seqüència de text crea una nova seqüència de text, un problema de gran interès científic. Ja que una gran quantitat de problemes presents al camp de "Natural Language Processing"són generalitzables d'aquesta manera, com pot ser la traducció automàtica, que és transformar el text de la

seqüència original a la seqüència traduïda. Aquesta publicació va portar a la creació dels "Transformers", un nou tipus de sistema de deep learning, creat específicament pel problema de seqüència a seqüència. Aquests nous models poden entrenar molt més ràpid que models antics [31]. I degut a millores sobre la seva arquitectura, poden arribar a multiplicar la seva velocitat d'aprenentatge per factors superiors a 1.000 [30], això els permet aprendre sobre molta més informació en temps que abans podria trigar mesos. Això es extremadament important, ja que abans, la quantitat de textos utilitzats per aprendre era menor, per a mantenir el temps d'aprenentatge a nivells raonables. Ara es pot aprendre sobre terabytes de text amb facilitat, millorant la qualitat del model generat.

2.1.1 Conceptes lingüístics

El treball utilitza certs conceptes lingüístics i criteris per a la realització dels anàlisis, en concret s'utilitzen les categories gramaticals, les coreferències, la polisèmia, les distàncies sintàctiques, la llei de Zipf i la llei de la brevetat.

El primer concepte son les categories gramaticals. Aquestes son classificacions de les paraules agrupant paraules amb propietats gramàtiques similars, es a dir, paraules amb la mateixa categoria gramatical solen tenir funcions similars en la estructura sintàctica del text, per exemple, els adverbis o noms. Per aquest treball, s'utilitzen les categories gramaticals definides al Penn Treebank Project [1].

El segon concepte son les coreferències, aquestes son la referencia a un mateix objecte en un text, per exemple, en la frase 'En Pau menja, ell menja ràpidament', Pau i ell son una coreferència. Aquesta equivalència pot estar separada per una quantitat arbitràriament llarga de paraules en text, i la detecció automàtica de coreferències computacionalment es un problema encara d'una gran dificultat, fins i tot pot presentar un problema per a la detecció humana en alguns casos. No existeix una llei que dicti el comportament de les coreferències, però si que segueixen una tendència, i es que un objecte no es referenciat moltes vegades en un mateix text. La tendència es que en general, els objectes en un text varien, ja que només hi ha tant que es pugui mencionar sobre un objecte.

El tercer concepte es la polisèmia. La polisèmia es defineix com el nombre diferents de significats que te una paraula i ve representat per un conjunt de significats representat per *synsets* tal com estan definits al diccionari. Aquests synsets son un set de una o mes polisèmies, tal com a definides a WordNet [11], per exemple, la paraula 'ratolí' tindria de polisèmies ratolí d'ordinador i ratolí d'animal. Considerem la

cardinalitat d'aquest conjunt de synsets (o significats) de la paraula com el grau de polisèmia o el nombre de sentits diferents que te la paraula. Aquest significant la mesura de la generalització expressada per aquesta paraula, es a dir, termes més específics tindran de per mitja menys significats, i termes més generals, en tindran més.

El quart concepte son les distàncies entre dependències sintàctiques, aquesta es la distancia, en paraules, entre dues paraules sintàcticament relacionades, es a dir, que existeix una relació gramatical entre les paraules. No existeix ninguna llei que dicti el comportament de les distàncies sintàctiques, però si que s'observa en els humans una tendència a reduir les distancies, per tant, tendeixen a ser petites, ja que el context d'una conversació es pot perdre si aquesta distància es molt gran, com es observat a [19].

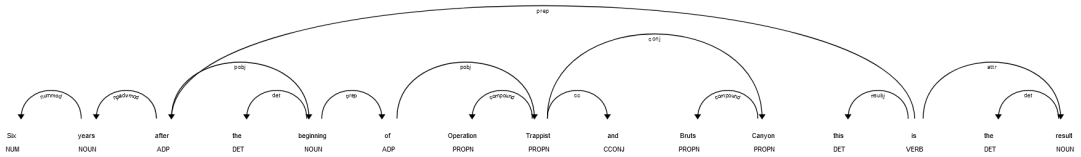


Figura 1: Exemple dependències sintàctiques sobre una frase, produït amb SpaCy [29]

El cinquè concepte, es la llei de Zipf [36], aquesta llei dicta la relació entre la freqüència d'una paraula (f) i la seva posició a una llista ordenada per freqüències (k), aquesta relació està dictada com a

$$f(k, s, N) = \frac{k^{-s}}{\sum_{n=1}^{N}(n^{-s})} \tag{1}$$

Essent N la mida del vocabulari, k la posició en la llista ordenada per les freqüències i s el exponent que caracteritza la distribució. Aquesta llei es present en tot llenguatge analitzat [35]. La llei de Zipf indica que el text humà no es aleatori i segueix una distribució, la qual podem modelar, a la vegada, la llei de Zipf només modela la distribució de paraules. Un text generat per un procés de zipf probablement no tindrà ninguna lògica. Encara així, al no ser un procés aleatori, podem observar si existeix una diferencia entre la distribució en humans i en el text maquina.

El sisè concepte es la llei de la brevetat [36], aquesta llei dicta que paraules més freqüentment tindran una mida més petita, mesurada aquesta mida en caràcters, i que existeix una relació negativa entre la mida de la paraula i la seva freqüència. Existeixen moltes explicacions per aquesta llei, la més freqüent sent la llei

del menor esforç, la llei que dicta que l'acció més comú serà aquesta que requereix de menys esforç. Al igual que la resta de lleis, ja que no es un procés aleatori, i que per a una màquina el principi del menor esforç no existeix, podem analitzar per buscar una diferencia.

El setè concepte es la llei de Hillberg, aquesta dicta com l'entropia condicional dels caràcters en un text redueix respecte a la quantitat de caràcters considerats per calcular el següent caràcter. Tal que H(n)=a*nˆB +hn, amb B≈0.5 i h≈0. Aquesta llei defineix que eventualment, l'entropia de tot text tendirà asimptoticament a una entropia condicional de 0, d'aquesta llei podem extreure que el text no es un procés aleatori, igual que es va concloure en zipf. Ja que el text humà no es un procés aleatori, el text màquina tindrà que imitar aquesta distribució per a ser similar al humà. La entropia condicional es definida com a:

$$\mathrm{H}(Y|X) = -\sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)} \tag{2}$$

2.2 Descripció del problema

Els models de llenguatge creats amb la arquitectura de xarxa neuronal transformer han sigut ràpidament provats en tot tipus de tasca de processament de llenguatge natural, i s'ha progressat fins crear el model GPT2 [27], aquest model excedeix en la generació automàtica de textos. Excedeix tant, que la versió completa del model no es va publicar fins al 5 de novembre de 2019 [28], amb la justificació de por al potencial per crear fake news que el model mostrava. Aquesta capacitat va primer ser trobada quan el model va crear una historia coherent sobre unicorns parlants.

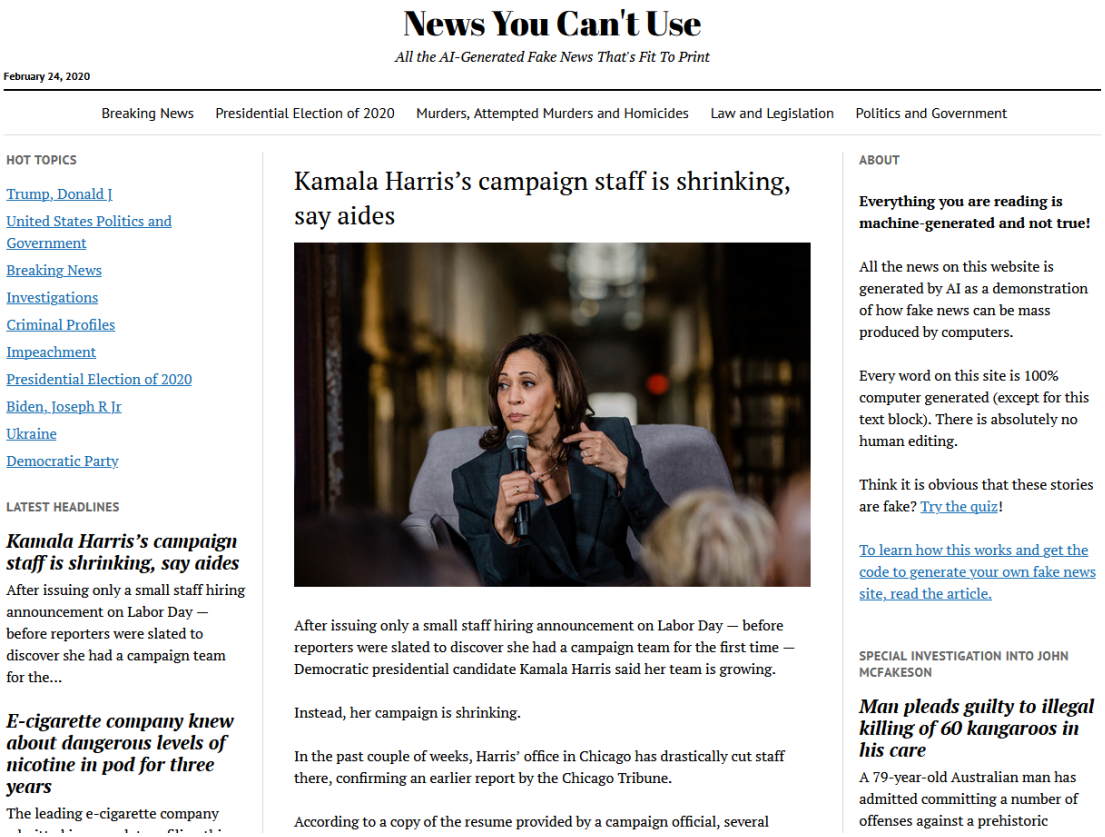
```
In a shocking finding, scientist discovered a herd of unicorns living in a remote,
    previously unexplored valley, in the Andes Mountains. Even more surprising to the
    researchers was the fact that the unicorns spoke perfect English.
```

Figura 2: Text Original [26]

```
The scientist named the population, after their distinctive horn, Ovids Unicorn. These four-
    horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally
    solved.

Dr. Jorge Prez, an evolutionary biologist from the University of La Paz, and several
    companions, were exploring the Andes Mountains when they found a small valley, with no
    other animals or humans. Prez noticed that the valley had what appeared to be a natural
    fountain, surrounded by two peaks of rock and silver snow.

 Prez  and the others then ventured further into the valley. By the time we reached the top
    of one peak, the water looked blue, with some crystals on top, said Prez.

 Prez  and his friends were astonished to see the unicorn herd. These creatures could be seen
     from the air without having to move too much to see them  they were so close they could
     touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also
    spoke some fairly regular English. Prez stated, We can see, for example, that they have
    a common language, something like a dialect or dialectic.

Dr.  Prez  believes that the unicorns may have originated in Argentina, where the animals
    were believed to be descendants of a lost race of people who lived there before the
    arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created
    when a human and a unicorn met each other in a time before human civilization. According
     to Prez, In South America, such incidents seem to be quite common.

However, Prez also pointed out that it is likely that the only way of knowing for sure if
    unicorns are indeed the descendants of a lost alien race is through DNA. But they seem
    to be able to communicate in English quite well, which I believe is a sign of evolution,
     or at least a change in social organization, said the scientist.
```

Figura 3: Text generat per GPT2 [26]

Com es pot veure al text mostrat a la figura 3, amb el text de la figura 2, pot generar articles sencers coherents, amb fortes implicacions negatives sobre el potencial d'aquest model, com és també mencionat a [28], on es parla sobre la capacitat de crear propaganda per supremacisme blanc, marxisme, jihadisme islàmic i anarquisme. Podem trobar exemples on aquest model és utilitzat per generar una pàgina que només conté fake news construïdes per aquest model, com podem observar a la Figura 12.



News You Can't Use

All the AI-Generated Fake News That's Fit To Print

February 24, 2020

Breaking News | Presidential Election of 2020 | Murders, Attempted Murders and Homicides | Law and Legislation | Politics and Government

HOT TOPICS

Trump, Donald J
United States Politics and Government
Breaking News
Investigations
Criminal Profiles
Impeachment
Presidential Election of 2020
Biden, Joseph R Jr
Ukraine
Democratic Party

LATEST HEADLINES

Kamala Harris's campaign staff is shrinking, say aides

After issuing only a small staff hiring announcement on Labor Day — before reporters were slated to discover she had a campaign team for the...

E-cigarette company knew about dangerous levels of nicotine in pod for three years

The leading e-cigarette company

Kamala Harris's campaign staff is shrinking, say aides

After issuing only a small staff hiring announcement on Labor Day — before reporters were slated to discover she had a campaign team for the first time — Democratic presidential candidate Kamala Harris said her team is growing.

Instead, her campaign is shrinking.

In the past couple of weeks, Harris' office in Chicago has drastically cut staff there, confirming an earlier report by the Chicago Tribune.

According to a copy of the resume provided by a campaign official, several

ABOUT

Everything you are reading is machine-generated and not true!

All the news on this website is generated by AI as a demonstration of how fake news can be mass produced by computers.

Every word on this site is 100% computer generated (except for this text block). There is absolutely no human editing.

Think it is obvious that these stories are fake? Try the quiz!

To learn how this works and get the code to generate your own fake news site, read the article.

SPECIAL INVESTIGATION INTO JOHN MCFAKESON

Man pleads guilty to illegal killing of 60 kangaroos in his care

A 79-year-old Australian man has admitted committing a number of offenses against a prehistoric

Figura 4: Pàgina de fake news completament artificial [23]

Mentre que la generació de text ha millorat exponencialment, encara no existeixen eines per ajudar a detectar aquests textos màquina. Al ser molt recents, aquests models generadors de llenguatge no han sigut encara explorats i per tant, errors amb aquests models, o punts forts, encara no són coneguts. Llavors, un sistema que pugui oferir informació sobre las característiques del text és d'alta necessitat per contrarestar usos il·legítims d'aquesta nova tecnologia, a la vegada que pot oferir un millor coneixement sobre els models generadors de text.

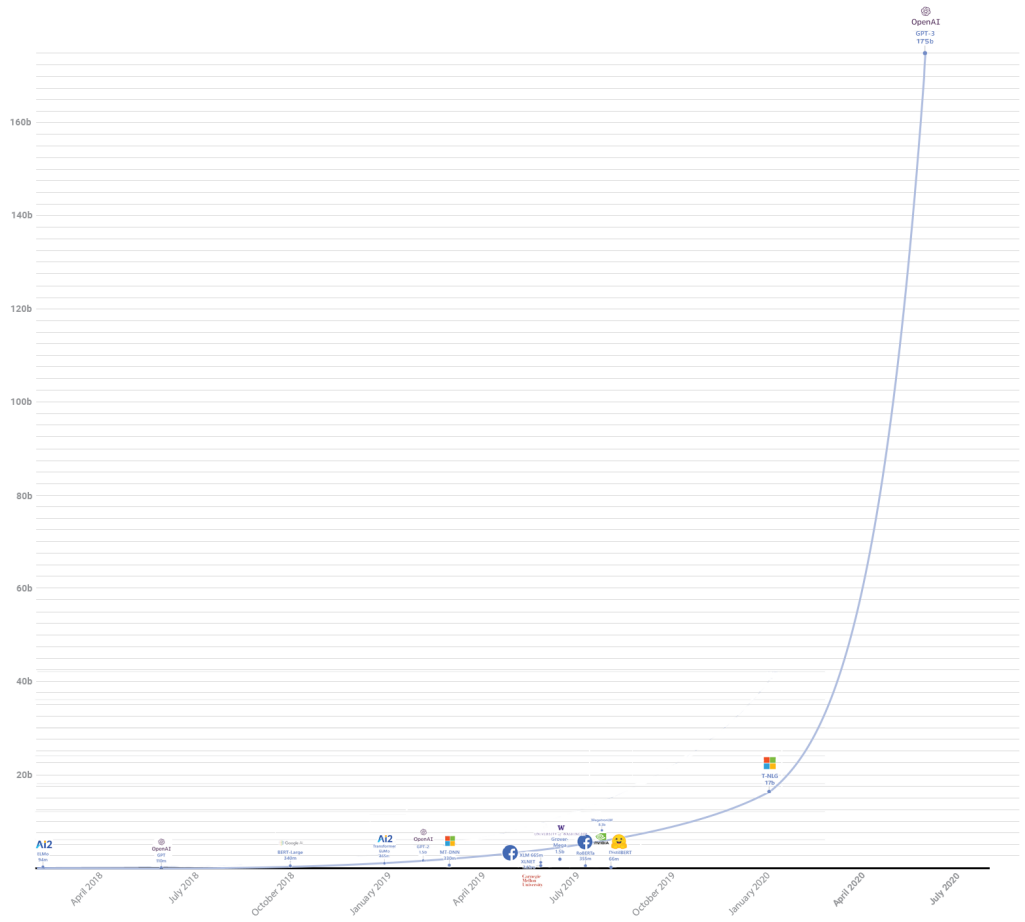


Figura 5: Creixement exponencial dels models transformer, en bilions de paràmetres del model [14]

Recentment, s'ha publicat el més potent d'aquests models, GPT3 [7], una millora directa sobre GPT2. Aquest model és superior a GPT2, però, amb 175.000 milions de paràmetres, requereix per sobre de 300GB en RAM utilitzant una representació reduïda, en coma flotant a 16 bits. Tals requeriments fan que usos il·legitims d'aquest model siguin extremadament improbables, a més, encara no existeixen datasets per analitzar aquest model, i la generació d'aquests es impossible amb els recursos presents en aquest treball. Encara així, els resultats obtinguts per GPT3 poden ser perillosos, i demostren la potencia d'aquesta tecnologia, ja que humans només han pogut tenir una precisió del 52% quant tenien que diferenciar noticies generades per GPT3 amb noticies humanes [7]. De fet, els autors de gpt3 no han fet públic el model gpt3 a temps d'ara.

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	"I don't know" assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

Figura 6: Precisió al diferenciar noticies escrites per humans i GPT3, tal com presentat a la figura 3.11 de [7]

2.3 Actors Implicats

Els actors implicats en el projecte poden ser dividits en els següents:

- **Director:** El director és una figura vital en el projecte, no només dirigeix aquest, sinó que ofereix coneixement expert per a la realització d'aquest, ja que aquest projecte tracta amb tecnologia innovadora.
- **Investigadors:** L'anàlisi de text generat per models de llenguatge com GPT2 és encara un camp sense explorar, i un que podria permetre entendre millor aquests models, per tant hi ha un interès científic sobre l'anàlisi d'aquest text.
- **Desenvolupadors:** El monitoratge de sistemes amb fort input dels usuaris, com poden ser per exemple, xarxes socials o pàgines de compra en línia és important per evitar atacs maliciosos, poder detectar text automàtic amb una certa precisió suposaria una eina interessant per desarrolladors d'aquests sistemes.
- **Usuaris de xarxes socials:** Ja que són el grup més exposat als possibles usos negatius d'aquests models generatius de llenguatge, són el grup que més es pot beneficiar una vegada s'implementés aquest classificador en aplicacions d'ús regular, o per exemple, com a plugins sobre un navegador web.

3 Abast del projecte

3.1 Solucions Existents

No existeix ningun software que extregui les propietats esmentades del text, ja que normalment si es necessari es fa en un cas individual per a cada programa. La extracció d'aquestes propietats només existeix, en alguns casos, en biblioteques de software que implementen mètodes per extreure-les, no en un software pur. Encara així, ja que el propòsit es oferir un software que ofereixi eines per a millor entendre el text, especialment per a fer més fàcil la classificació d'aquest en humà o no humà, considerarem les solucions actuals per diferenciar text.

Sent un problema de cert interès, ja existeixen sistemes classificadors que intenten diferenciar text. Específicament existeixen dues categories. Fora d'aquestes categories, no s'han creat encara discriminadors de text automàtic amb text humà, això es deu al fet que la creació de text d'aquesta qualitat és un fet nou, i abans la diferenciació es considerava una tasca trivial.

La primera categoria se centra en la mateixa tecnologia que GPT2, que són classificadors de seqüència en l'arquitectura RoBERTa [20], un transformer. Aquests classificadors però, només han sigut entrenats per discriminar un model generador de text, i no generalitzen el seu coneixement, encara així, mostren bona precisió contra el model amb el qual són entrenats.

La segona categoria, es centra en utilitzar el propi model per discriminar, on es mira com de probable és cada text en el model, i es classifica respecte a aquesta probabilitat, com més probable que vingui del model, més probable que sigui generat. Avui en dia només existeix una solució així [13]

Així, les solucions existents no són suficients per abordar el problema de discriminació de text, ja que o requereixen preguntar al model propi, o ser entrenades en un model propi. Això porta a una generalització de la informació extremadament pobre, ja que una solució generadora de text basada en regles, podria fàcilment passar un filtre utilitzant la segona categoria. Per tant és necessària una solució que utilitzi informació existent al text que no requereixi coneixement del model.

3.2 Justificació de la solució escollida

Per trobar la solució que permeti adreçar els problemes amb les solucions existents, és necessari explorar propietats especials del llenguatge, que un model màquina no pugui aprendre, o tingui gran dificultat per reproduir-les. Especialment, utilitzant el camp de la lingüística quantitativa, on existeixen lleis sobre el llenguatge que permeten classificar i diferenciar de text automàtic.

D'aquestes lleis, són de gran interès la llei de Zipf i la llei de Hillberg, la llei de Hillberg dictant l'entropia condicional d'un text, i la llei de Zipf dictant la freqüència de les paraules. Mentre que la llei de Hillberg encara no s'ha utilitzat per a aquests fins, la llei de Zipf sí que s'ha utilitzat abans per a classificació automàtica de textos [34].

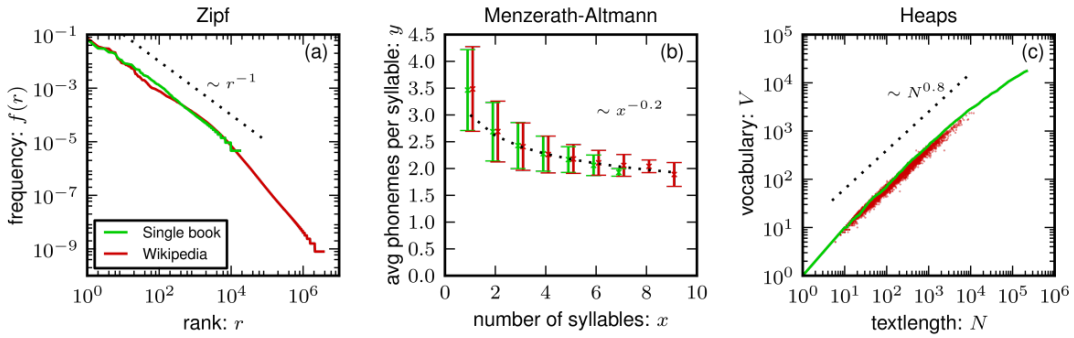


Figura 7: Diferents gràfics representant diferents lleis de la lingüística quantitativa [5]

Per tant, la solució proposta és, utilitzant lleis de la lingüística quantitativa, especialment la llei de Zipf i la llei de Hillberg, caracteritzar els textos, per a així oferir característiques d'aquests textos per a ajudar la creació de eines més potents per diferenciar el text.

3.3 Objectius

El projecte és format per un objectiu principal, amb dos sub-objectius. L'objectiu principal és obtenir informació dels textos que ajudi a caracteritzar millor el text i el seu origen. Això es farà amb un software que extraurà les propietats del text, però perquè aquest software pugui funcionar, requerirà dos sub-objectius.

- **Anàlisi de text automàtic** Per poder comprendre més apropiadament el nou paradigma en la generació de text, es farà un anàlisi sobre el text automàtic respecte a les lleis de la lingüística quantitativa. D'aquesta manera es pot arribar a comprendre millor les seves propietats, i d'aquestes

discriminar entre els textos, ja que propietats que no ajudin a diferenciar entre text humà o màquina no ajudarien al propòsit del software.

- **Extracció de les propietats** Finalment, és necessari escollir aquelles propietats que diferencien el text humà de la màquina, i crear un software que extregui aquestes. Aquest software haurà de ser fàcil d'utilitzat, i simple d'integrar en pipelines de processament de dades per aprenentatge màquina, ja que principalment serà utilitzat per enriquir les dades que arribarien a un algoritme d'aprenentatge màquina.

3.4 Riscos i obstacles

El projecte té, com a obstacles principals, 3 riscos que poden manifestar-se durant el desenvolupament d'aquest.

- **Similitud entre text automàtic i humà**

  Es casi impossible, abans d'analitzar a fons els textos, veure si les lleis de la lingüística quantitativa es compleixen o no en els textos automàtics. Per tant, existeix un fort risc de que no existeix una propietat, entre les analitzades, que permeti diferenciar text només amb l'ús de lingüística quantitativa. En aquest cas el software agafarà un focus amb textos generats no per el state of the art, si no per software més antic, que si que son diferenciables per aquestes característiques.

- **Baix rendiment**

  L'extracció d'informació del text podria portar a algoritmes d'extremadament baix rendiment, i, encara que la informació obtinguda sigui important, si el temps d'execució es excessiu, no seria atractiu per utilitzar en situacions de producció.

- **Temps insuficient**

  L'anàlisi de textos i la creació del software analitzador és un procés lent, i podria arribar a la situació on una falta de temps porta a escollir una solució subòptima, ja que analitzar apropiadament els valors necessaris podria requerir uns mesos extres.

3.5 Metodologia i rigor

El projecte seguirà una metodologia àgil, amb reunions setmanals amb el director del projecte, Mario Martin Muñoz, cada dimarts, per a valorar el progrés i establir les tasques a fer, amb un fort èmfasis en la creació de classificadors operatius. Ja que execucions per analitzar valors, o provar classificadors poden necessitar de dies a setmanes, les tasques no seran centrades de setmana a setmana, sinó amb un temps variant depenent de la mateixa tasca.

La metodologia àgil és escollida, ja que poc és conegut sobre els texts i les seves qualitats, i una solució en cascada no podria tenir en compte la forta variabilitat dels possibles requisits que es manifestin, ja que pot ser necessari adaptar el classificador per utilitzar noves lleis lingüístiques, o altres canvis que requereixen una forta flexibilitat que no seria present a una metodologia clàssica de desenvolupament de software.

3.5.1 Eines

Per al compliment de la metodologia esmentada abans, es farà ús de diferents eines, la primera serà per al control de versions del codi on s'utilitzarà Git, amb una branca remota de seguiment a Github. Per al control de les tasques, s'utilitzarà Trello, que permet el seguiment de les tasques d'una forma àgil. El codi on el projecte serà escrit es Python 3, gràcies al fort suport que ofereix per a processament del llenguatge natural i per a la creació de classificadors, cosa que permetrà accelerar moltes tasques relacionades al preprocessament de text. Per a executar l'anàlisi i entrenar els classificadors, fa falta una forta màquina, per això es requerirà un servidor personal amb recursos suficients per fer execucions que poden durar setmanes.

4 Obtenció dels textos

Els textos han sigut obtinguts del projecte de github de Open-AI anomenat gpt-2-output-dataset [24], en aquest repositori son disponibles diferents textos, d'aquests es van obtenir els que tenien menys input humà i utilitzaven el model GPT2 més potent, el xl-1542M, aquests estàn format per 250.000 textos generats per GPT2 XL-1542M. Acompanyant aquests textos son els webtext, que estan formats per també 250.000 textos, però aquests d'origen humà. En general, els textos humans son més petits que els textos de GPT2, per això s'han normalitzat les dades en certes distribucions.

S'ha escollit aquests textos ja que generar nous seria extremadament car en recursos i temps, ja que GPT2 es un model gran, especialment el model XL-1542M, que està composat per 1.542 milions de paràmetres. A la vegada, l'ús de dades en accés públic i dels creadors de el propi model afavoreix a l'estandardització dels anàlisis. Permetent comparació de resultats amb major facilitat, ja que el treball estableix un nivell base de classificació de textos creats per GPT2, aquest valor es més precís si es compara en les mateixes dades.

Per a fer els anàlisis s'han agafat els textos anteriors i s'han fusionat en blocs de 25.000, mantenint si eren humans o de GPT2 respectament, i s'ha fet l'anàlisi sobre aquests blocs de text. Aquesta fusió s'ha fet degut a la alta variància en les distribucions que seria observada en blocs de textos mes petits, per això s'ha fusionat el text, per reduir aquesta variància, permitent un anàlisis més precís. Això també significa que mentre que els resultats observats son repetibles amb textos suficientment llargs, no ho son amb petits textos, com els de les dades utilitzades sense fusionar. Encara així, marquen les tendències que les dues distribucions segueixen.

```
What the team was looking for, in particular, were elements like thorium and uranium, which
    along with potassium, warm Earth's interior. This heat affects its plate tectonics and,
    according to the scientists, the way it retains its water. Though the functions of that
    heat-to-plate-to-liquid interaction aren't fully understood -- it's "one of the great
    mysteries in the geosciences," the study's advisor, Wendy Panero, put it -- scientists
    have speculated that the forces of heat convection in the mantle, the ones that move
    Earth's crust, have some kind of role in regulating the amount of water in the oceans. "
    It seems that if a planet is to retain an ocean over geologic timescales, it needs some
    kind of crust 'recycling system,' and for us that's mantle convection," Unterborn said.
Which means, in turn, that plate tectonics could also be a key indicator of a planet's
    hospitality to life. Particularly for microbial life -- since microbial life on Earth,
    the study's authors point out, benefits from subsurface heat. (Take the single-celled
    microbe archaea, some of which live not off the energy of the sun, but rather off the
    heat rising from inside the Earth.) And that indicator, the team reasoned, can be
    approximated by analyzing a given exoplanet's sun: the more thorium in the star, say,
    the more likely a terrestrial planet formed around that star would be to support life.
    Since it would stand to reason that the planets that orbit around those suns contain
    more thorium, as well, that would suggest that the interiors of those exoplanets are
    warmer than ours -- and also that those planets are more geologically active than Earth.
     And that would mean that they are more likely than Earth to retain the liquid water
    that supports life.
And: Of the eight solar twins the team studied, seven of them seemed to contain more thorium
     than our own star.
One star in the team's survey, for example, contained 2.5 times more thorium than our sun.
    And per the study's measurements, terrestrial planets that formed around that star
    likely generate 25 percent more internal heat than Earth does -- with all that that
    implies. So there could be at least one planet that is potentially more life-affirming
    than Earth. But there could also be, the study suggests, more where that came from:
    exoplanets that are more earthly than Earth -- planets nourished and made hospitable by
    the warmth of other suns.
Now that the hunt for exoplanets has moved from science fiction to science, that finding --
    preliminary and tentative as it is -- could have implications for humans' ability to
    find signs of life elsewhere in the universe. As study co-author Cayman Unterborn summed
     it up: "If it turns out that these planets are warmer than we previously thought, then
    we can effectively increase the size of the habitable zone around these stars by pushing
     the habitable zone farther from the host star, and consider more of those planets
    hospitable to microbial life."
```

Figura 8: Text exemple de webtext

```
Six years after the beginning of Operation Trappist and Bruts Canyon this is the result:
The highest quality, heirloom, wild caravel vineyards in the world.
Brasenose valleys are named after the Mediterranean mint plant, Eylea petunia, which has
    long been used to make a bitter liqueur among the Ancient Romans. The background was
    transferred by Paul Riboud (a savant/baker) from his roots in spectacular Moravia and
    early Christian traditions. The incredibly tender berry of in-season production from
    this magnificent vineyard, Boulla, hides the bitterness of the aged grape and
    compliments Chocolate Milk.
These wines are excellent raisins, sultanas, of delicious full flavor, vibrant redrocks and
    meadows of slightly verdant madrone and vervain tannins; all exuding a "we're-going-to-
    have-too-much-beer" vibe!
Innovation and artisanship have combined in some astounding wines from the lowest of primary
     soils: "The Corner", a ruby mineral black that eventually turns to scarlet, and "Beaver
     Bay", a deep ebony, occupying an entire hillside rock facing evocative of an Old World
    binaural bubble bath.
Keep the wine off the food! Leave it to the practicioners in these wonderful wine regions.
Read more about us on The Traveller by the Pyany Hotel and returned to the original post for
     original discussion.
It is important to note that these wines contain a great deal of politically incorrect
    information and an equal willingness to give bootlegged ebooks as dishes. Suddenly
    Taking With Skin la conference Nention Sitje hosted in Lesbos last week has must have
    had hundreds of people attending where he also gave "cultural training" about the stuff.
     One of the presentations on the topic of food was on learning how to deal with "
    unnecessary exploitation" etc. Whos's Getting That Dream Now which by Emily Esfahani
    Smith used as our title the word skimpy to describe foods in sour landscapes.Also whoaa
    got one was the sound of water trickling down mountainsides since water can be shared
    from lift tickets. No wonder some of the food was in a dialect of Croatian!Reciprocation
     at its finest.One day I will take this serious agro-socially and REALLY import and
    import LOTS of fruit on the rivers and the great green leaves of the islander oak trees
    (which look like windswept legumes) and make bargains with local families and apartments
     who will run the bins in exchange for my fruit.It's just a matter of town or state
    where you purchased alliances!
```

Figura 9: Text exemple de GPT2 XL-1542

5 Anàlisi dels textos

Per a solucionar el problema de poder diferenciar text, primer es necessari transformar-ho a un nou espai on pugui ser tractat computacionalment. Per fer això, la única manera es extreure informació del text, però aquesta informació ha de tenir una propietat, i es que respecte a la generació de text humà, aquesta propietat no es aleatòria, es a dir, o es modelable, o segueix una evolució clara. S'escolleixen aquestes característiques ja que ofereix informació característica del text humà, si un text intentés replicar aquestes qualitats, tindria també que replicar aquestes característiques, però això ha sigut en el passat de una dificultat intractable. Encara així millores recents en la generació de text podrien crear característiques similars al text humà.

A la vegada, mentre que el text humà es relativament simple de diferenciar del text màquina per l'ull humà, això es degut a que el humà te un model de llenguatge que es capaç de comprendre i extreure el significat darrere del text, per exemple, la frase 'El heiwheo ahir era interessant' no te el menor sentit per a un humà, però una màquina no pot veure aquesta diferencia fàcilment. Per això, han de ser característiques que una màquina pot extreure.

Per a la elecció de la informació a extreure, s'han escollit múltiples propietats que el text humà conté, algunes d'elles modelables amb distribucions estadístiques, altres per tendències generals.

5.1 Anàlisi respecte a lleis cognitives

5.1.1 Anàlisi de la proporció de categories gramaticals

Com ha sigut mencionat anteriorment, les categories gramaticals son classificacions de les paraules agrupant paraules amb propietats gramàtiques similars. Aquestes han sigut calculades amb la biblioteca de python, nltk, amb les categories gramaticals presents sent les presents al Penn Treebank Project [1]. Després han sigut agregades per fer l'anàlisi.

La proporció de categories gramaticals es manté relativament semblant entre el text humà i el text de GPT2, la diferencia més significativa entre les dues distribucions es del 0.6%, encara així existeixen diferencies, i aquestes diferencies son molt més grans que la variància de les distribucions, per tant, son suficients per a diferenciar perfectament entre text de GPT2 o humà.

Malgrat tot ha de destacar-se la quantitat necessària de text per poder tenir aquesta distinció. Quan en experiments s'han tractat textos individuals, no s'ha tingut aquesta diferenciació tant clar. Sent la variància molt superior a los diferencies trobades.

La diferencia més clara és amb NNP (Noun Proper, Singular , es a dir, un nom que es refereix a un objecte en concret, com pot ser Lluna), de 0.006. Els NNPs són aproximadament un 12% de les paraules, per tant farien falta un mínim de 167 NNP, o 1392 paraules aproximadament per veure una diferencia d'un sol NNP entre les dues distribucions. Per tant caldria un text de més de 69600 paraules com a mínim per veure una diferencia, només en aproximadament 50 paraules d'aquest text.

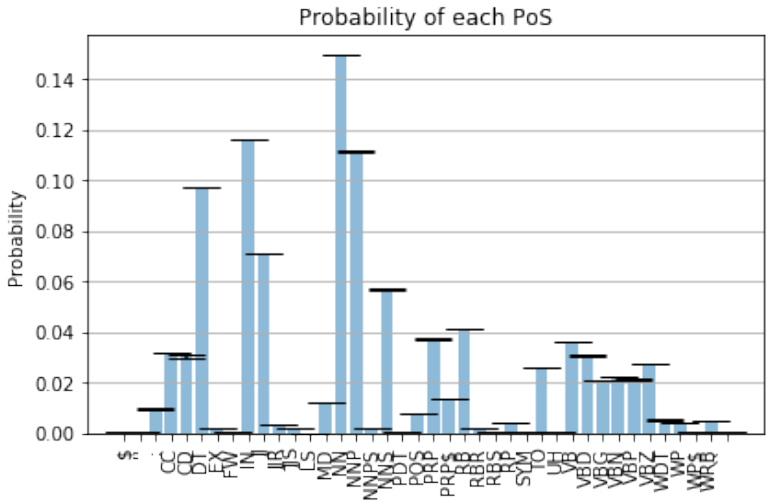


Figura 10: Proporció categories gramaticals de text humà.

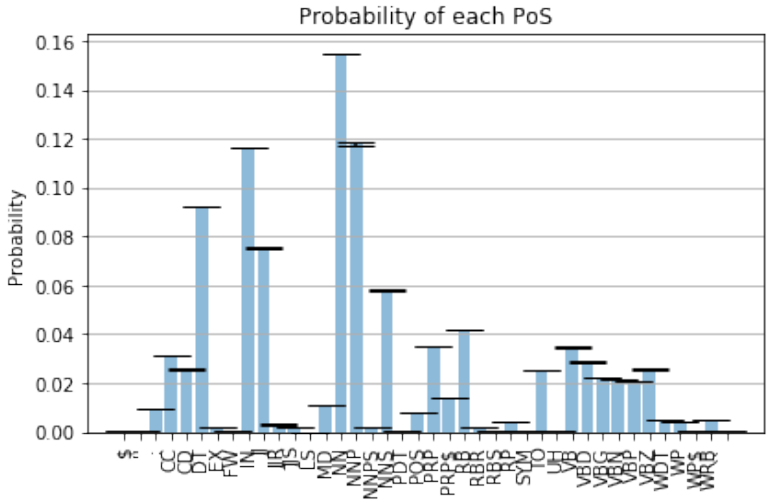


Figura 11: Proporció categories gramaticals de GPT2.

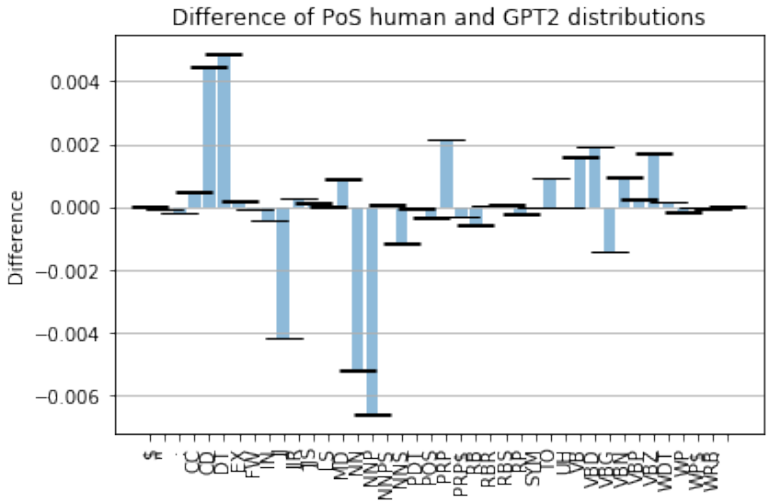


Figura 12: Diferencia de proporcions de categories gramaticals entres text humà i GPT2

5.1.2 Hillberg's law

Com ha sigut mencionat anteriorment, la llei de Hillberg [15] dicta com l'entropia condicional dels caràcters en un text redueix respecte a la quantitat de caràcters considerats per calcular el següent caràcter. Tal que H(n)=a*nˆB +hn, amb B≈0.5 i h≈0.

S'analitza la distribució de l'entropia condicional en tres aspectes del text, les paraules del text, les categories gramaticals del text, i els caràcters del text. Encara que la llei de Hillberg només ha sigut descrita sobre caràcters, podem utilitzar el concepte en altres aspectes del text. Això es per a observar si el model te un coneixement sobre la distribució en tots els nivells, ja que podria ser que sap crear una distribució apropiada en les paraules, però pobre en les categories gramaticals.

5.1.2.1 Text pur

No hi ha diferencia observable entre GPT2 i el text humà entre les figures 13 i 14, excepte en que GPT2 hi ha una major variància. Encara així, existeix una clara diferencia entre les distribucions present al text humà i a GPT2, com podem observar a la figura 16, això indica que, mentre que les distribucions son diferents, aquesta diferencia només pot ser observada apropiadament si la variància es suficientment reduïda. Si la variància es suficientment menor, separar entre text generat per humans o GPT2 es trivial, com es observable a la figura 16, on els textos humans formen un grup separat dels textos de GPT2 i viceversa. També s'observa que la llei de Hillberg no segueix la forma característica que ofereix en els caràcters, però si que segueix una forma clarament definida i diferent.

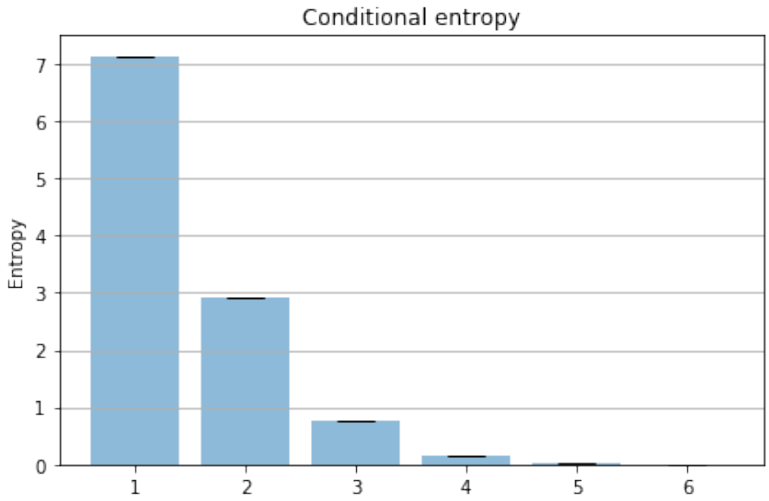


Figura 13: Entropia condicional del text pur humà

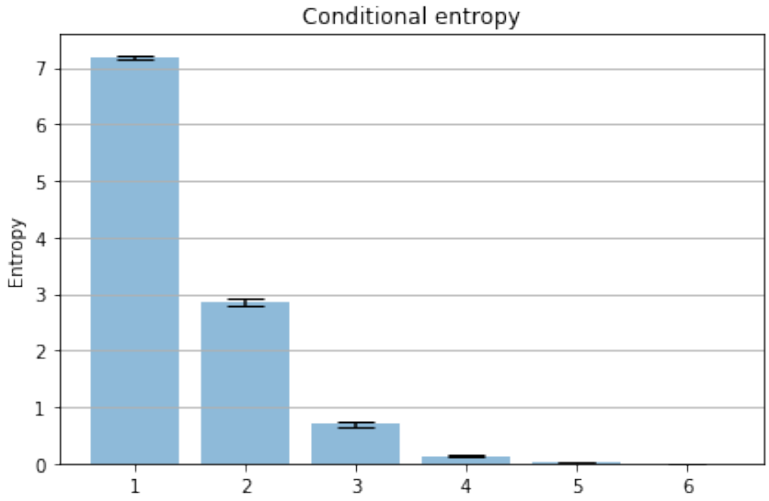


Figura 14: Entropia condicional del text pur GPT2

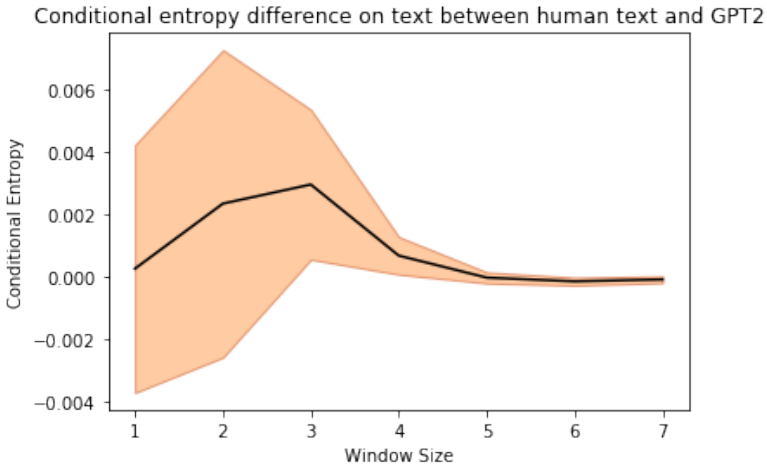


Figura 15: Diferencia entre distribucions de entropia condicional sobre text pur entre text humà i GPT2

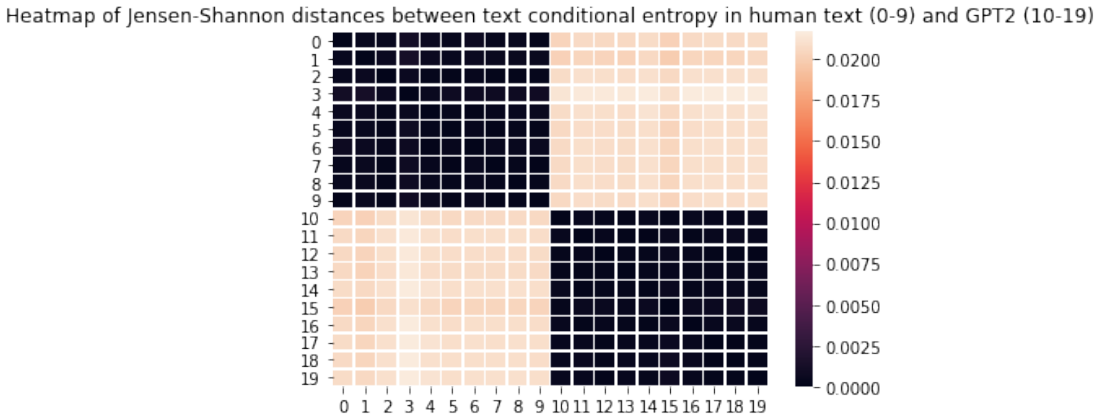


Figura 16: Divergència Jensen-Shannon [18] entre distribucions de Hillberg sobre text pur en text humà i GPT2

5.1.2.2 Categories gramaticals

Com abans, no hi ha diferencia observable entre GPT2 i el text humà entre les figures 17 i 18, excepte en que GPT2 te una major variància. Però, exactament com abans existeix una clara diferencia entre les distribucions present al text humà, com observat a 20 i a GPT2, i es deu a els mateixos fenòmens als observats anteriorment, la diferència, mentre que encara es significativa, es menor, això indica que el model te una forta similitud en la distribució de categories gramaticals comparat amb el text humà, i que distribueix les estructures sintàctiques de manera molt similar a la humana. Al igual que amb el la distribució de entropia condicional en les paraules, també s'observa que la llei de Hillberg no segueix la forma característica que ofereix en els caràcters, però si que segueix una forma clarament definida i diferent.

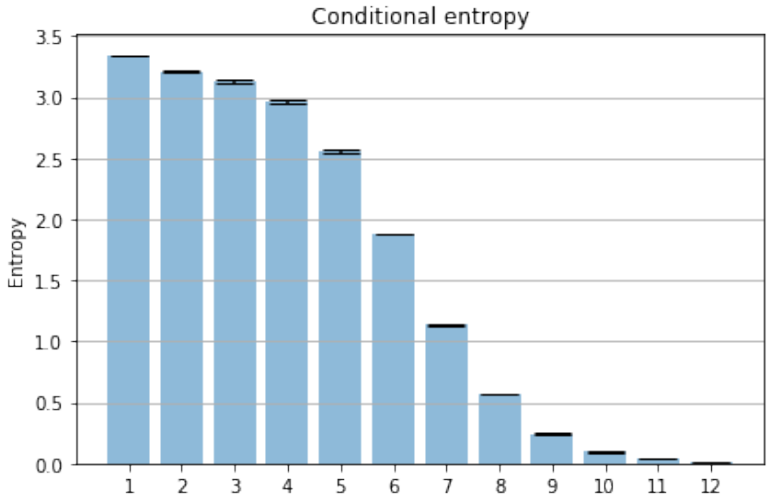


Figura 17: Entropia condicional de les categories gramaticals del text humà

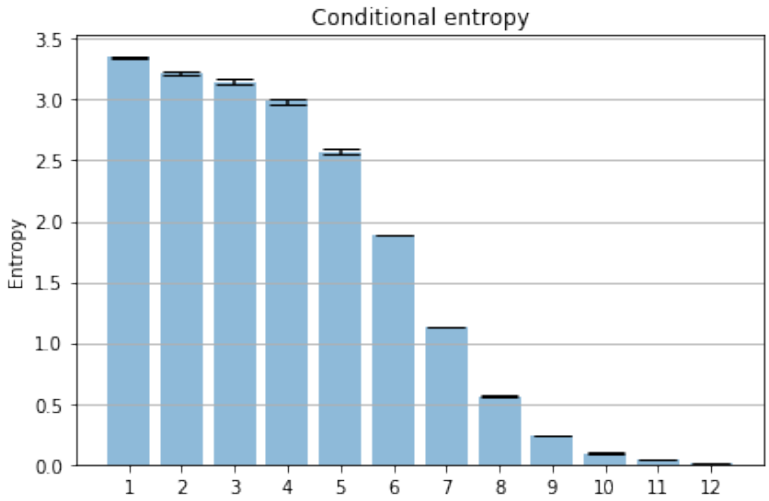


Figura 18: Entropia condicional de les categories gramaticals del text GPT2

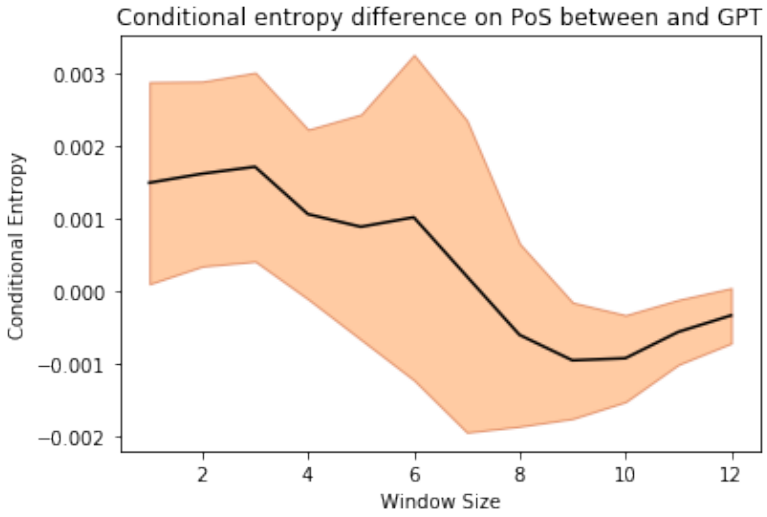


Figura 19: Diferencia entre distribucions de entropia condicional sobre categories gramaticals entre text humà i GPT2

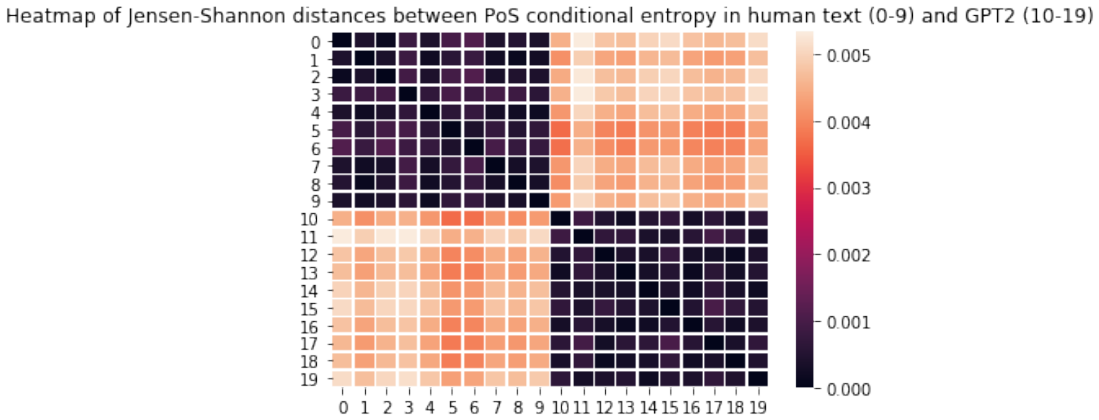


Figura 20: Divergència Jensen-Shannon [18] entre distribucions de Hillberg sobre categories gramaticals en text humà i GPT2

5.1.2.3 Caràcters

Al igual que en les categories gramaticals, i el text pur, no hi ha una diferencia evident a ull entre les figures 21 i 22 , encara així, podem observar algunes propietats interessants, la primera, es que la variància sobre les distribucions es menor en el rang de 4 a 5 caràcters, que coincideix amb mides comunes de paraules, però observem una major variància fora d'aquest rang, com podem veure a la figura 23, encara així, com

hem observat abans, aquesta es minúscula en termes absoluts, però suficientment significant com per a diferenciar trivialment entre GPT2 i humà, com podem veure a la figura 24. D'interès es que la entropia de GPT2 es major fins arribar a la mida de finestra de 8 caràcters, on el text humà comença a sobrepassar en entropia.

La major diferencia observada entre les dues distribucions es de, com a molt, 0.0010 en la figura 23, molt menor a la variància possible, encara així, com podem veure a la figura 19, la distància entre distribucions es suficientment significativa com per a diferenciar entre textos de GPT2 i humans trivialment.

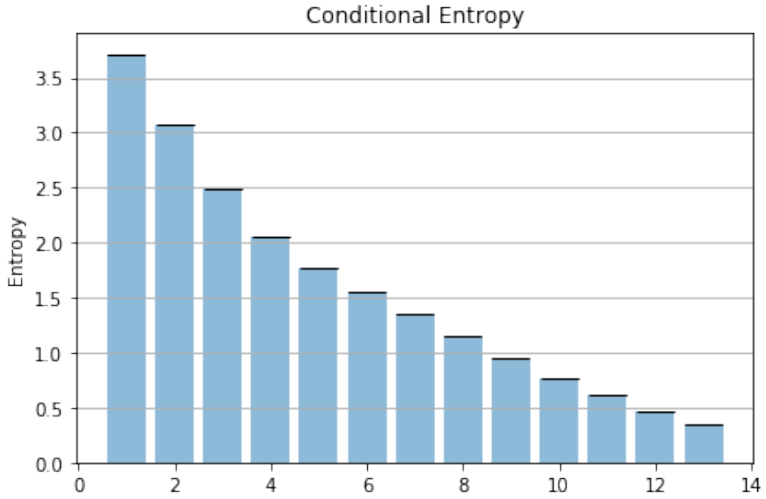


Figura 21: Entropia condicional de els caràcters del text humà

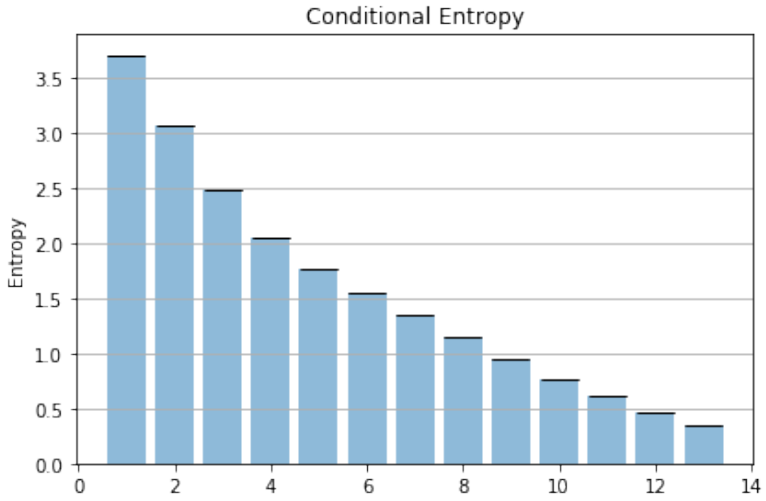


Figura 22: Entropia condicional de els caràcters del text humà del text GPT2

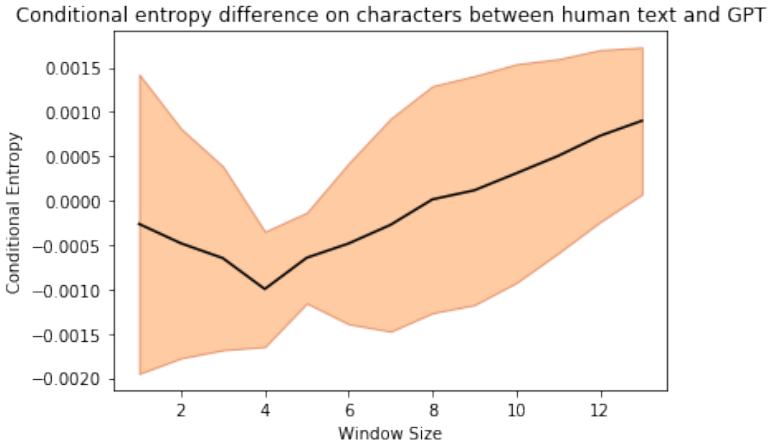


Figura 23: Diferencia entre distribucions de entropia condicional sobre caràcters entre text humà i GPT2

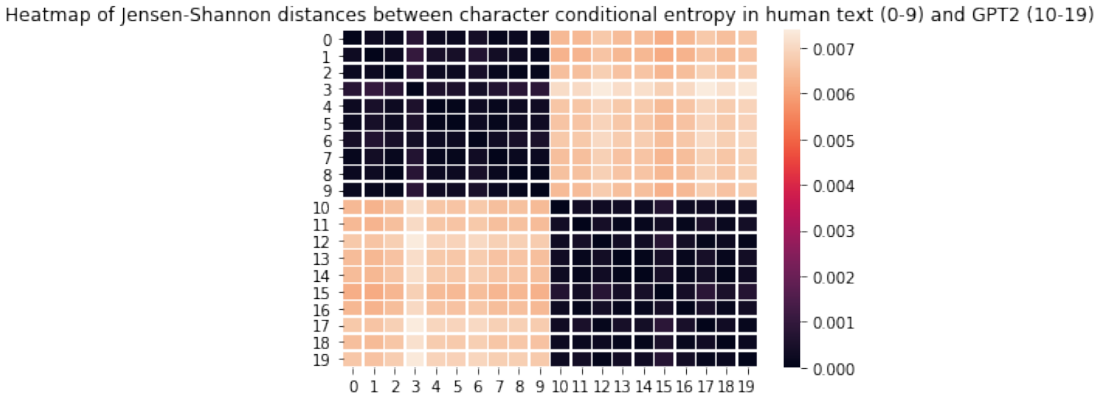


Figura 24: Divergència Jensen-Shannon [18] entre distribucions de Hillberg sobre caràcters en text humà i GPT2

5.1.3 Zipf

Com ha sigut mencionat anteriorment, la llei de Zipf es una llei que dicta la relació entre la freqüència d'una paraula (f) i la seva posició a una llista ordenada per freqüències (k), aquesta relació està dictada com a

$$f(k, s, N) = \frac{k^{-s}}{\sum_{n=1}^{N}(n^{-s})} \tag{3}$$

Essent N la mida del vocabulari, k la posició en la llista ordenada per les freqüències i s el exponent que caracteritza la distribució.

No hi ha diferencia observable entre GPT2 i el text humà entre les figures 25 i 26 excepte que GPT2 utilitza un major vocabulari, si fitem les dues distribucions a una distribució de zipf, trobem que per al text humà, $s = 1.0289$ i per a GPT2 $s = 1.0291$. Per tant les dues distribucions son casi completament idèntiques, cosa que indica que la llei de zipf es una característica poc útil per diferenciar entre textos. Malgrat tot, es pot observar en l'ajust de la llei que aquesta es marcada per dos règims [12], i un anàlisi més profund en aquests règims podria trobar alguna diferencia no observada en aquest treball.

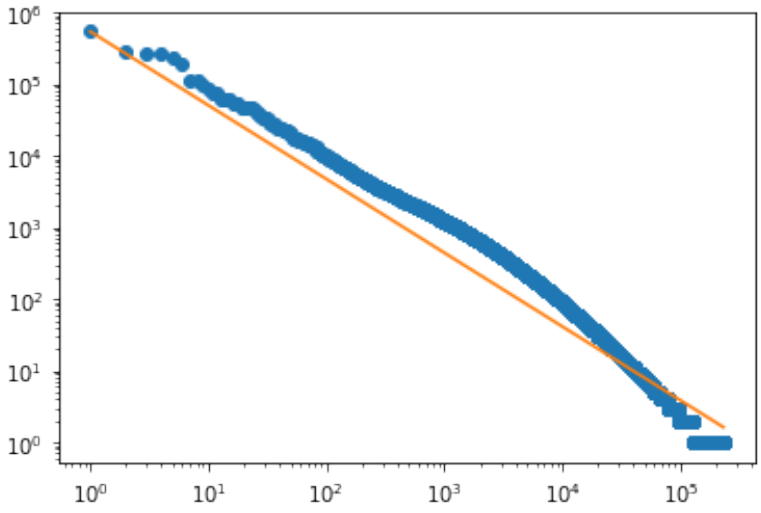


Figura 25: Freqüència paraules en text humà

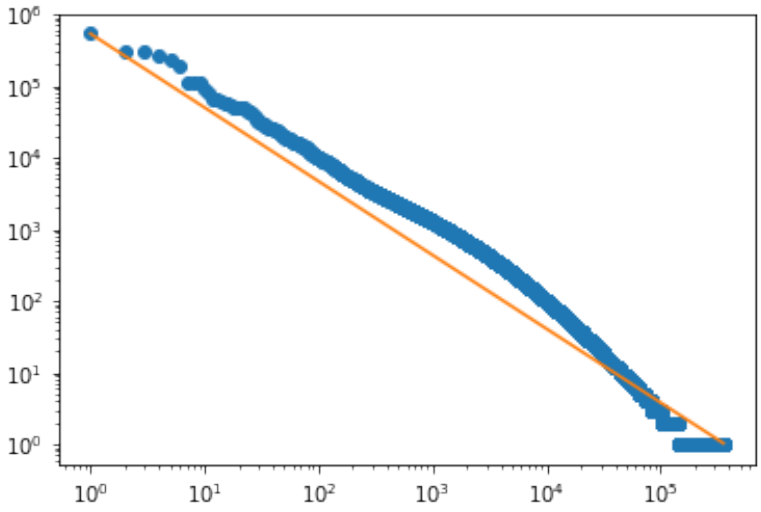


Figura 26: Freqüència paraules en text GPT2

5.1.4 Probabilitat mida paraula

Com ha sigut mencionat anteriorment, la llei de la brevetat dicta que paraules més freqüentment tindran una mida més petita, mesurada aquesta mida en caràcters, i que existeix una relació negativa entre la mida de la paraula i la seva freqüència. Per mesurar aquesta llei, s'han agregat totes les paraules amb mides iguals, i s'ha generat una distribució estadística amb els resultats.

Les distribucions generades han calculat la probabilitat, per a cada mida de paraula possible observada, la seva probabilitat de ser observada. No hi ha una gran diferencia observable entre GPT2 i el text humà, com observat a les figures 27 i 28 excepte que GPT2 utilitza paraules més llargues, però podem fixar-nos en algunes característiques interessants.

La primera, es la quantitat de paraules amb mides no presents al vocabulari humà, això es per la presencia de brossa a els textos, com, en el cas de els textos humans, la existència de paraules com 'rawdownloadcloneembedreportprint', mentre que a GPT2 podem trobar paraules com 'BitmapIgnoreHostToScreenRequestTokenToBlockDrawablesWithTimeout'. Això indica que es possible filtrar GPT2 observant brossa, però les probabilitats son extremadament baixes de trobar-ne aquesta, en els textos observats de GPT2, només un 0.10% dels textos tenien paraules d'una mida superior a 80 caràcters, per tant, com a molt, podria filtrar un 0.10% de textos generats per GPT2.

La segona, es la diferencia de fins a un 0.6% de el text humà respecta a GPT2, observada en la figura 29, quant x=3, que indica que el text humà utilitza, de per mitja, més paraules de mida petita que GPT2, en el costat contrari, veiem que GPT2 utilitza moltes més paraules llargues que el text humà, això seguiria la idea que el model ignora la mida d'una paraula per a escollir-la. Encara així, les diferencies son menors al 1%.

Analitzant només les paraules de mida de dos caràcters, de probabilitat aproximadament 15%, amb una diferencia del 0.6% en ambdues distribucions, faria falta al menys 167 paraules de mida de tres caràcters per veure una diferència d'una sola paraula, per tant amb 1114 paraules, podríem esperar una diferencia d'una sola paraula. Per a 50 paraules de diferencia en mida tres, necessitaríem aproximadament unes 55700 paraules.

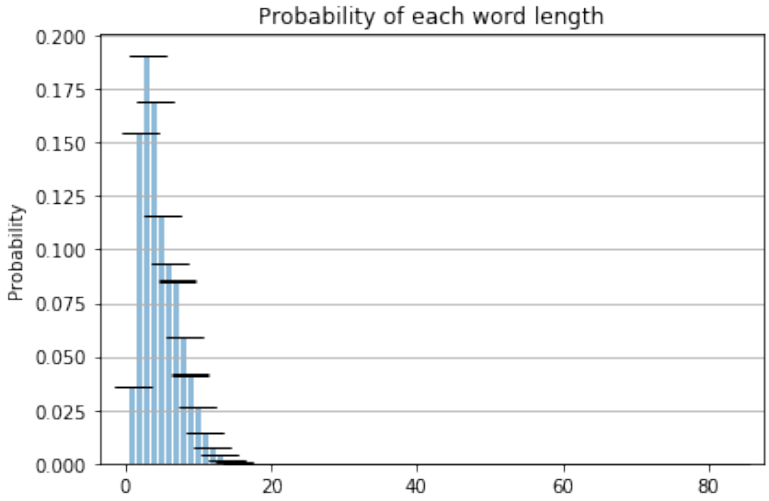


Figura 27: Distribució mida paraula en text humà

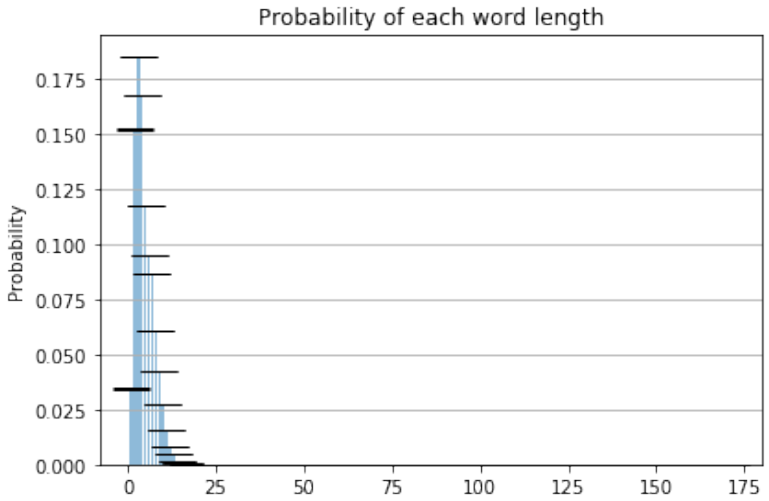


Figura 28: Distribució mida paraula en text GPT2

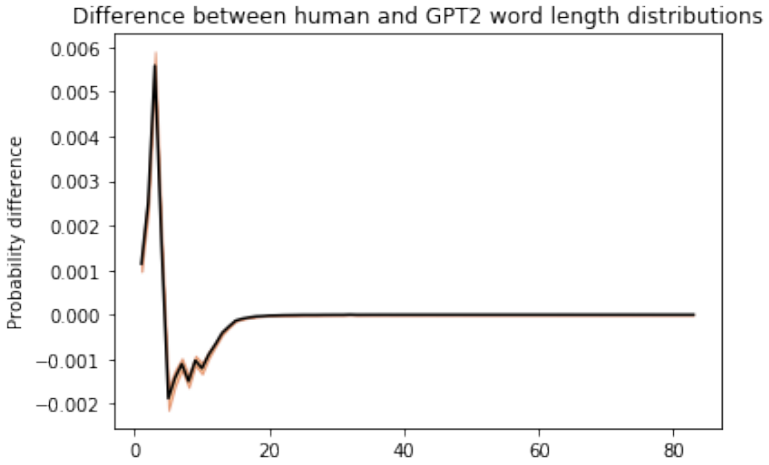


Figura 29: Diferencia distribució mida paraules entre text humà i GPT2

5.2 Anàlisi de dades semàntiques

5.2.1 Distancies de dependències sintàctiques

Com ha sigut mencionat anteriorment, definim la distància de dependència sintàctica com a la distancia en el nombre de paraules que separa dues paraules que presenten una dependència gramatical, per exemple, en 'Jo avui menjo', hi ha una distància sintàctica de 1 entre jo i menjo.

Analitzem la distribució de la probabilitat de cada distancia entre dependències, aquesta dependència obtinguda per SpaCy [29]. Com podem veure a la figura 32, el text humà te mes dependències semàntiques petites, exceptuant les distàncies de mida 1, mentre que GPT2 en te més de llargues, fins a arribar un punt on la diferencia es insignificant, a més, la variància es molt menor a la diferencia, per tant, existeix una clara diferencia entre GPT 2 i el text humà en aquesta distribució, encara així, aquesta només es visible quant la variància es suficientment petita, es a dir, quant el test es suficientment gran.

La major distancia es quant la distancia es tres, en un 0.4%, amb aquesta distancia tenint una probabilitat aproximada del 10%, per tant faria falta analitzar un mínim de 2500 distancies aproximadament per obtenir una diferencia d'una sola distancia entre les dues distribucions. Per a veure una diferencia de 50 distàncies de mida 3, farien falta com a mínim 125.000 distancies.

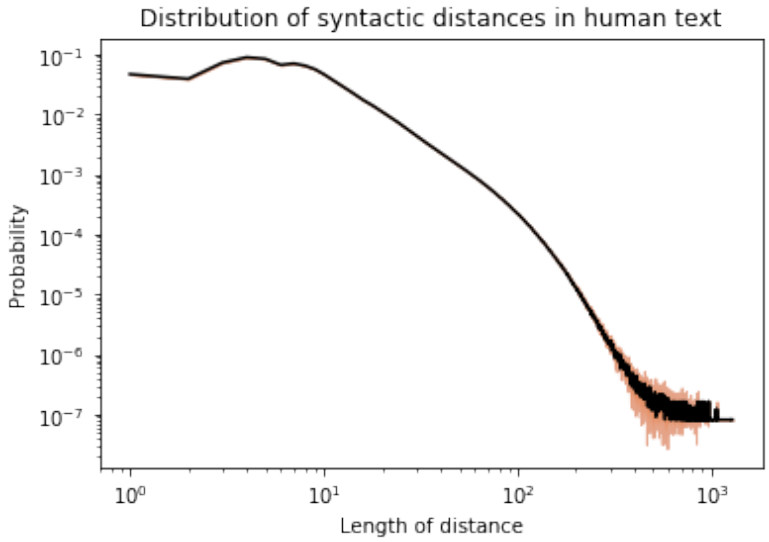


Figura 30: Freqüència de distancies sintàctiques del text humà

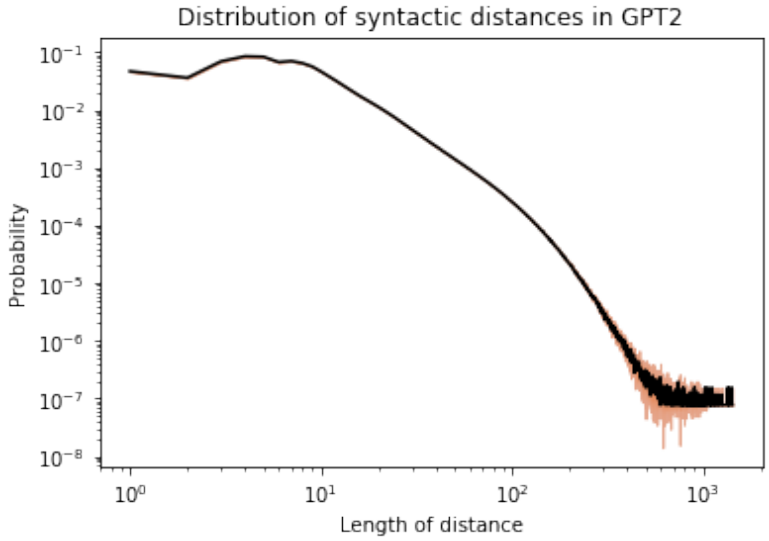


Figura 31: Freqüència de distancies sintàctiques del text GPT2

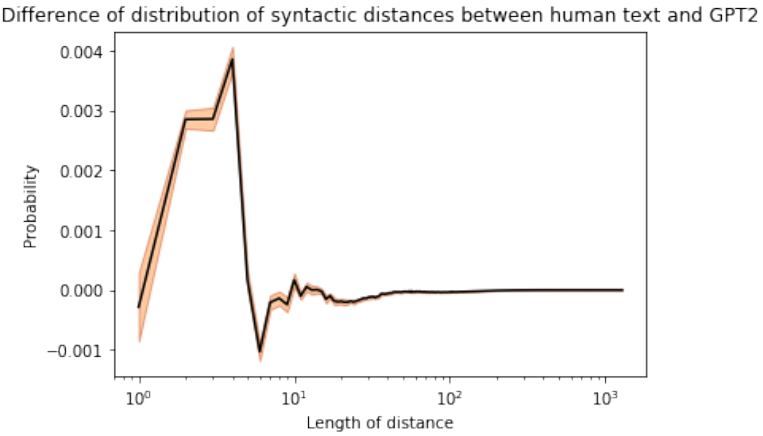


Figura 32: Diferencia entre distribucions de distàncies sintàctiques entre text humà i GPT2

5.2.2 Polisèmia

Com ha sigut mencionat anteriorment, la polisèmia es defineix com el nombre diferents de significats que te una paraula. Definim el grau de polisèmia com el nombre de synsets existent que no son la pròpia paraula, es a dir, una polisèmia de zero vol dir que la paraula només te un significat.

Analitzant la distribució de la probabilitat del nombre de synsets de cada paraula no existeix una diferencia significant entre la figura 33 i la figura 34, en els dos textos manté una forma extremadament semblant, encara així, la variància entre les distribucions es tan baixa que permet diferenciar entre textos humans i GPT2 amb perfecta precisió sobre aquests textos analitzats. D'interès especial, es com GPT2 te aparentment mes paraules amb menys synsets comparat amb el text humà, encara si les diferencies de probabilitat son del 0.3% com a molt, com observat a 35.

Mentre que la major diferencia es en paraules amb una polisèmia de 14 significats, son exponencialment menys freqüents a les paraules amb una sola polisèmia, per tant, per veure una diferència d'un sol valor entre les dues distribucions, agafant les paraules amb una polisèmia, que tenen una probabilitat aproximada del 10%, amb una diferencia entre distribucions de aproximadament 0.2%, faria falta observar com a mínim 5000 paraules per trobar una sola diferencia. Per a observar 50 diferències farien falta 250.000 paraules.

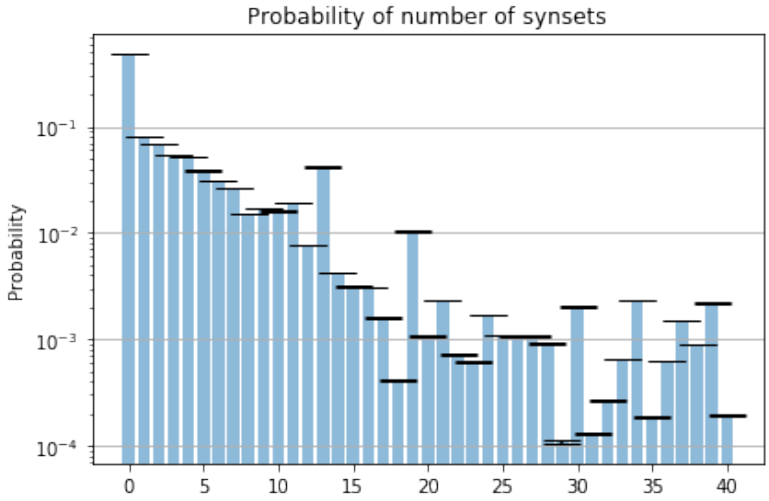


Figura 33: Distribució de nombre de synsets del text humà

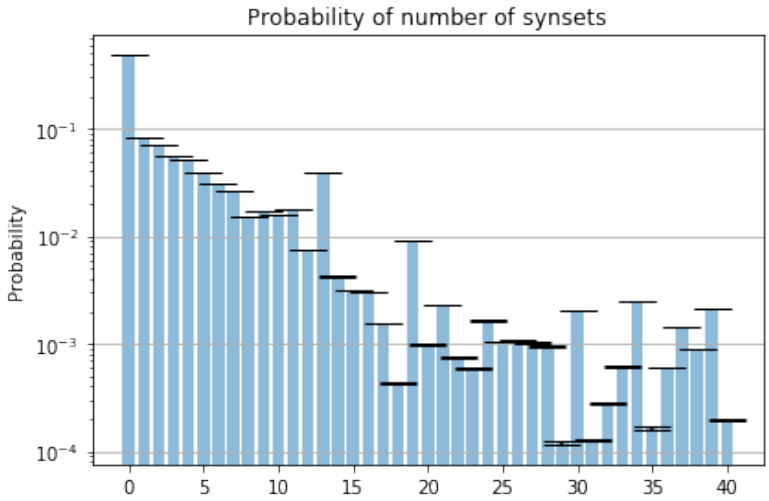


Figura 34: Distribució de nombre de synsets del text artificial

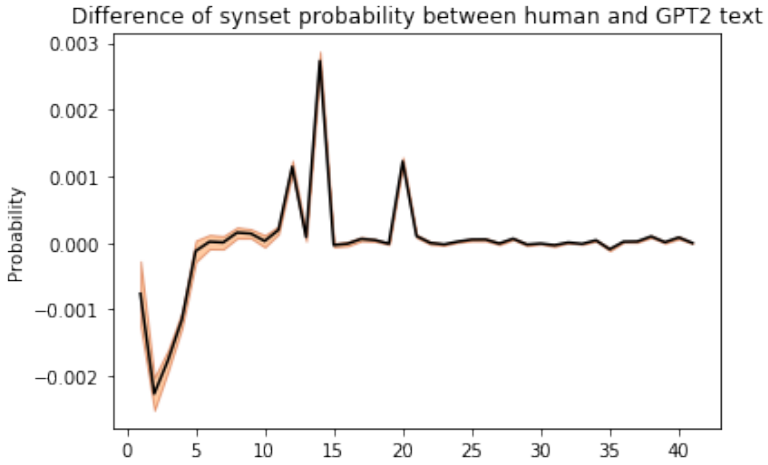


Figura 35: Diferencia de distribució de nombre de synsets entre text humà i GPT2

5.2.3 Coreferències

Com ha sigut mencionat anteriorment, aquestes son la referencia a un mateix objecte en un text, per exemple, en la frase 'En Pau menja, ell menja ràpidament', Pau i ell son una coreferència.

Per a trobar aquestes coreferències, s'ha utilitzat el state of the art existent, neuralcoref, de HuggingFace, tal com descrit a [33], encara així, errors son possibles, i algunes coreferències probablement son errònies, però amb la quantitat de text analitzat, l'efecte tindria que ser mínim, i afecta al text humà i GPT2 per igual.

A ull no es pot observar ninguna diferencia entre la figura 36 i 37, excepte en distàncies amb probabilitats rondant el 0.01%. Però al contrari, es poden veure diferències significatives en la figura 38, on es pot veure que hi ha una diferència de fins al 3.4% en un punt de la distribució.

Amb una diferencia entre distribucions d'aproximadament el 3.4% en grups de coreferències de mida 1, amb una probabilitat aproximada del 65% , per a veure una diferència d'un valor entre les dues distribucions en els grups de mida 1, farien falta aproximadament com a mínim 5 grups de coreferències, i per veure una diferencia de 50 valors, farien falta 250 grups de coreferències.

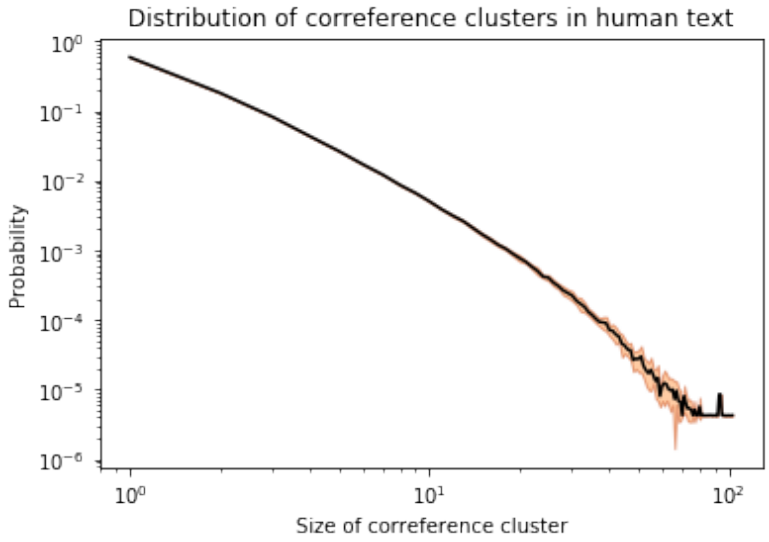


Figura 36: Probabilitat de clusters de coreferències del text humà

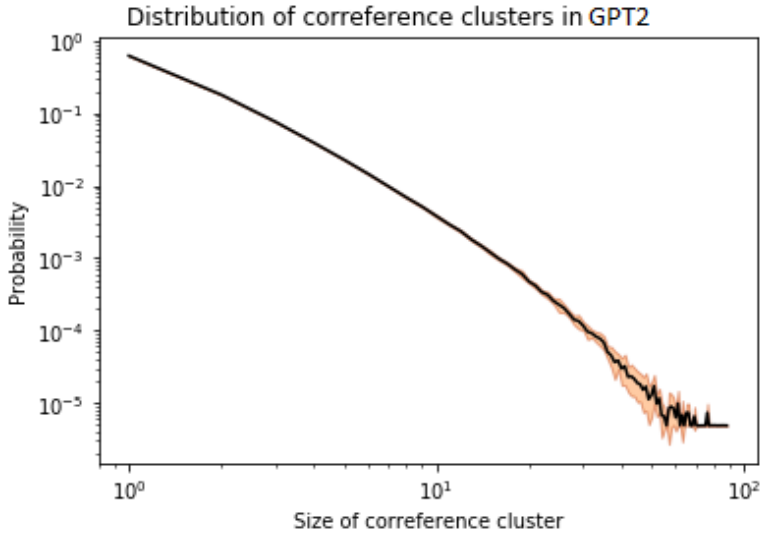


Figura 37: Probabilitat de clusters de coreferències del text GPT2

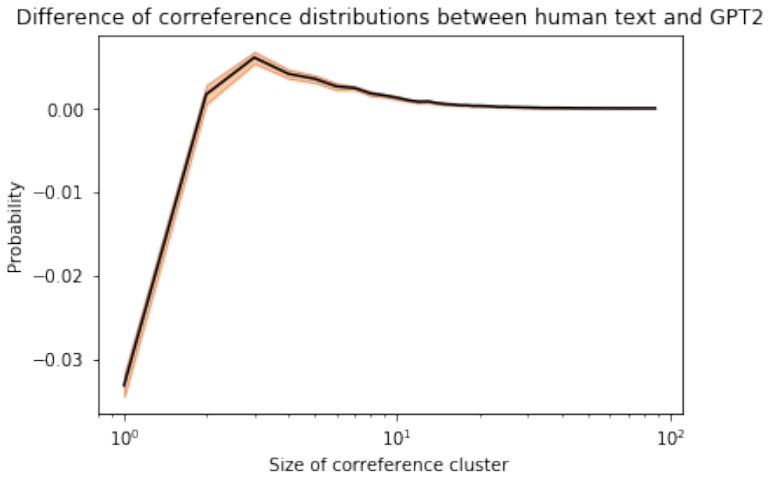


Figura 38: Diferencia de distribucions de coreferències entre el text humà i GPT2

6 Precisió de classificació de textos

La precisió s'ha calculat amb simple regressió logística [16] sobre els textos individuals, els 500.000 textos, 250.000 sent textos fets per GPT2 i 250.000 sent textos generats per humans i amb els texts fusionats, sent 10 textos de GPT2 i 10 textos humans.

Nombre de Textos fusionats	1	250000
Synsets	55.5%	100%
Distancies Sintàctiques	57.5%	100%
Categories gramaticals	60%	100%
Probabilitat mida paraules	57.7%	100%
Coreferències	55.5%	100%
Llei de Zipf	53.5%	100%
Llei de Hillberg Text	65.5%	100%
Llei de Hillberg Categoria Gramatical	55.5%	100%
Llei de Hillberg Caràcters	71.5%	100%
Fusió de característiques	74%	100%

Taula 1: Precisió al classificar text entre humà o GPT2

En conclusió, quant els textos son petits, diferenciar entre text màquina i humà es difícil utilitzant una sola propietat, encara així, existeixen diferencies entre els dos tipus de text, amb la més gran expressada per els valors de Hillberg sobre caràcters, seguit per la distribució de Hillberg sobre text pur, i finalment Hillberg sobre categories gramaticals, d'interès es que en Hillberg per categories gramaticals, la precisió es menor, així que mentre que la distribució es diferent, com hem vist a la distribució de categories gramaticals, la entropia condicional generada es similar, així que el model té una comprensió de l'entropia millor de les categories gramaticals que de la distribució d'aquestes.

Fusionant les característiques anteriors amb el mateix mètode de regressió logística [16], només s'arriba fins a un 74% de precisió al separar text màquina d'humà, sent una elecció de mètode relativament simple, son uns resultats interessants, i amb investigació amb solucions d'aprenentatge màquina, es podrien aconseguir resultats molt més prometedors.

Si incrementem la mida dels textos, la variància de les distribucions disminueix, i la petita diferencia entre les distribucions GPT2 i les humanes comença a ser mes i més fàcil de distingir, fins al punt on diferenciar entre els dos es relativament trivial.

7 Conclusions del anàlisis

Mentre que en agregat, els textos mostren una diferencia clara si son humans, o creats per GPT2, al nivell individual, els textos varien fortament, i fa detecció en textos, especialment textos petits, casi impossible amb les mètriques observades. Encara així, anàlisi a ull dels textos mostra que els textos de GPT2 no son tan coherents, o que parlen de coses que no poden ocórrer a la vida real, com el problema dels unicorns parlants. Això vol dir que fa falta una comprensió del llenguatge més profunda per a diferenciar els textos, una que pugui entendre la coherència de els textos, i a la vegada, que pugui fins i tot diferenciar per context, tal que fets impossibles en el mon real, però possibles per exemple, en un llibre, no siguen artificials.

Encara així, existeix una informació en aquestes mètriques, que mentre que per si soles no son suficients per diferenciar els textos, donen informació extra, que augmentada per altres mètodes si que podria arribar al nivell necessari per discriminar amb suficient precisió.

8 Disseny Software

Amb les propietats observades, podem crear un software que ofereixi aquestes mètriques per a així potenciar possibles mètodes que intentin classificar text, oferint-les com a metadades sobre aquests textos. Les metadades oferides seran aquestes analitzades anteriorment, que mentre que no ofereixen una gran quantitat d'informació, ofereixen informació i poden potenciar nous mètodes.

8.1 Arquitectura

Per a poder complir el requisit de fàcil integració, la arquitectura ha de oferir una interfície de programació d'aplicacions estàndard i d'ús extès, per això s'escolleix REST. Amb aquest disseny, la integració es agnòstica del llenguatge, permetent la integració a qualsevol sistema, i la possibilitat de desplegar la interfície de programació en un servidor dedicat per a poder així oferir una major escalabilitat a el sistema en el que s'integri.

Degut a la simplicitat del software propi, fora de ser una interfície de programació d'aplicacions, no conté més lògica, no requereix de base de dades, ni de complexa arquitectura, només requereix d'accés a biblioteques específiques de Natural Language Processing, encara així, les dependències específiques necessàries per a l'utilització d'aquestes biblioteques compliquen el disseny per evitar conflictes.

8.2 Documentació

Per a facilitar la integració fàcil, s'utilitza un sistema de documentació d'interfícies de programació REST, anomenat Swagger 2 [6], ja que aquest sistema de documentació es estàndard i extés, a la vegada que ofereix una simple interfícies gràfica web per a facilitar la comprensió de l'interfície de programació.

8.3 Implementació

El projecte es troba implementat a https://github.com/AntoniCasasM/TFG-TextAnalysis/tree/master

8.3.1 Llei de Hillberg

Per a la obtenció de les distribucions de Hillberg, s'ha implementat un algoritme per extreure la entropia condicional, definida com a:

$$\mathrm{H}(Y|X) = -\sum_{x\in X, y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)} \tag{4}$$

L'algoritme es simplement una iteració sobre el text, on x es la seqüència de mida equivalent a la finestra observada, i y es la paraula següent a aquesta. La iteració del vocabulari s'ha implementat com definit a l'algoritme 1, i l'extracció de l'entropia de les distribucions, s'ha implementat com definit a l'algoritme 2. El text varia depenent de l'anàlisi a fer, sent text pur, categories gramaticals o caràcters, depenent de l'anàlisis.

Algorithm 1 Obtenció de distribucions estadístiques

```
procedure OBTENTION OF DISTRIBUTIONS(text, windows) distributionsX distributionsXY
    for window ∈ windows do
     x = new Map                                   ▷ Inicialitzat tot valor nou a 0
     xy = new Map
     for val ∈ range(0, len(text) − window do
         for iter ∈ range(val, val + window) do
             sequence+=text[iter]
         x[sequence]+=1 xy[sequence][val+window]+=1
     normalize(x)
     normalize(xy)
     distributionsX[window]=x
     distributionsXY[window]=xy
```

Algorithm 2 Obtenció de entropia

```
procedure OBTENTION OF ENTROPY(distributionsX, distributionsXY) entropyDist
    for distributionX,distributionXY ∈distributionsX,distributionsXY do
   entropy=0
    for key ∈ distributionX do
        px=distributionX[key]
        for secondKey ∈ distributionX do
            pxy=distributionXY[key][secondKey]
            entropy+=-pxy*log2(pxy/px)        ▷ H(Y|X) = −∑ p(x,y) ∗ log(p(x,y)/p(x))
    entropyDist.append(entropy)
```

8.3.2 Llei de Zipf

Per a la implementació de la llei de Zipf s'ha utilitzat un algoritme simple que fa una iteració sobre el text per obtenir la distribució de Zipf, guardant els freqüències de cada paraula. Per facilitar el seu ús e integració, s'ha utilitzat una classe de python que ja implementa aquest tipus de algoritme, anomenada Counter.

Algorithm 3 Algoritme de Zipf

```
procedure ZIPF(text)
    zipfDistribution = new Map
    for value ∈ text do
     if ∈ then
         zipfDistribution[value]+=1
     else if ∉ then
         zipfDistribution[value]=1
  sort(zipfDistribution.values)
```

8.3.3 Coreferències

Per a la implementació de les coreferències, s'ha utilitzat el state of the art, una xarxa neuronal que soluciona coreferències, anomenada neuralcoref, de OpenAI [33], s'ha escollit així ja que es un problema increïblement complex. Mentre que això afecta a la mantenibilitat del sistema es l'única opció que permet obtenir suficient precisió per fer l'anàlisi.

8.3.4 Mida de paraula

Per a la obtenció de les distribucions de la mida de paraula, s'ha implementat el mateix algoritme amb unes petites modificacions, especialment la normalització de les dades per obtenir probabilitats, no freqüències. A la vegada, ja no s'ha utilitzat la classe de python3, Counter, ja que no es pot utilitzar per implementar aquestes modificacions.

Algorithm 4 Algoritme de Zipf modificat

```
procedure MODIFIED ZIPF(text)
    lengthDistribution = new Map
    for value ∈ text do
     value = length(value)                          ▷ Mida en caràcters
     if ∈ then
         lengthDistribution[val]+=1
     else if ∉ then
         lengthDistribution[value]=1
  normalize(lengthDistribution.values)
```

8.3.5 Polisèmia

Per a la obtenció dels synsets s'ha utilitzat l'enllaç que nltk [22] ofereix a WordNet [11], d'aquesta manera, obtenint la categoria gramatical utilitzant nltk per a especificar la paraula en el seu ús, i la base de dades de WordNet s'extreu la distribució de synsets als textos..

8.3.6 Distancies sintàctiques

La obtenció de les distancies sintàctiques s'ha obtingut amb spaCy [29], ja que entrenar un model per a poder obtenir aquests valors seria extremadament complex, s'ha escollit spaCy per fer aquesta tasca. La obtenció de les distancies es fa oferint les frases dels textos com a input a spaCy i restant entre la posició del pare del token, i la del token propi.

9 Planificació Temporal

El projecte requereix 540 hores, i té de duració de 10 de febrer al 2 de juny, és a dir, una duració d'unes setze setmanes. S'escull el 2 de juny com a data final, ja que no es pot conèixer el dia de presentació, per tant s'escull el dia més proper possible.

9.1 Definició de les tasques

En l'apartat següent es definiran les tasques pertinents al treball de final de grau. A cada tasca apareixerà una lleugera descripció junt amb el seu esforç en hores estimat. Si una tasca té dependències, s'indicarà.

9.1.1 Gestió del projecte

Reunions: T1

A causa de la metodologia àgil seguida, es farà setmanalment una reunió per avaluar els resultats obtinguts i prendre decisions necessàries respecte a aquests resultats obtinguts.

Temps estimat: 16h / 1h setmana

Recursos humans: Administrador del projecte

Recursos materials: Ordinador

Context i abast del projecte: T2

La contextualització del projecte i la definició de l'abast són una part necessària per poder entendre millor el problema a treballar, i el que fer, específicament perquè s'ha escollit una solució per sobre d'un altre, a qui afecta això i com es desenvoluparà aquesta solució.

Temps estimat: 25h

Recursos humans: Administrador del projecte

Recursos materials: Ordinador

Planificació Temporal: T3

Per a la gestió correcta del temps, és necessari fer una planificació temporal que permteixi establir estimacions per a l'esforç necessari per a cada tasca, i les seves fites temporals, d'aquesta manera, es pot seguir el progrés del treball amb més facilitat, i es poden fer canvis amb major facilitat.

Temps estimat: 8h

Recursos humans: Administrador del projecte

Recursos materials: Ordinador

Pressupost i sostenibilitat: T4

Una vegada feta la planificació temporal, és necessari fer un estudi de sostenibilitat i del pressupost necessari per al projecte, això permet contextualitzar el treball d'una altra manera, i permet veure diferents aspectes d'aquest, específicament permet contextualitzar l'efecte social, i el cost que suposaria. Temps estimat: 9h

Recursos humans: Administrador del projecte

Recursos materials: Ordinador

Elaboració de la memòria: T5

La redacció del document principal, que documenta el treball de final de grau, aquest document és necessari per a la finalització del mateix TFG.

Temps estimat: 70h

Recursos humans: Administrador del projecte

Recursos materials: Ordinador

Presentació: T6

Una vegada finalitzada la memòria, s'haurà de preparar una presentació sobre el treball de final de grau i els seus resultats obtinguts.

Temps estimat: 5h

Dependències: T5

Recursos humans: Administrador del projecte

Recursos materials: Ordinador

9.1.2 Anàlisi

Anàlisi de la llei de Zipf: T7

S'ha de realitzar un anàlisi sobre la llei de Zipf [17], la llei que dicta la relació entre freqüència d'una paraula, i posició en una llista ordenada per freqüències, específicament com els textos artificials segueixen aquest comparat amb els textos humans, centrant-se en com divergeixen.

Temps estimat: 24h

Recursos humans: Lingüista computacional

Recursos materials: Ordinador, Servidor

Anàlisi de la llei de Hillberg en text: T8

S'ha de realitzar un anàlisi sobre la llei de Hillberg [15], que dicta la relació entre l'entropia condicional de les paraules, específicament com aquesta tendeix a reduir-se quan més paraules es considerin abans de la següent paraula, s'analitzaran els textos artificials i els textos humans centrant-se en com divergeixen.

Temps estimat: 25h

Recursos humans: Lingüista computacional

Recursos materials: Ordinador, Servidor

Anàlisi de la llei de Hillberg en categories gramaticals: T9

Semblant a l'anàlisi de la llei de Hillberg amb text pur, s'ha de realitzar un anàlisi amb la llei de Hillberg sobre els constituents, això és per analitzar si existeix una divergència entre els models maquina respecte als humans.

Temps estimat: 25h

Recursos humans: Lingüista computacional

Recursos materials: Ordinador, Servidor

Anàlisi de la llei de Hillberg en caràcters: T10

Semblant a l'anàlisi amb lleis de Hillberg anteriors, s'ha de realitzar un anàlisi amb la llei de Hillberg sobre els caràcters del text, això és per analitzar si existeix una divergència entre els models maquina respecte als humans.

Temps estimat: 25h

Recursos humans: Lingüista computacional

Recursos materials: Ordinador, Servidor

Anàlisi de coreferències: T11

Com esmentat per Open-AI [25], el text generat té una major quantitat de pronoms, això fa un anàlisi sobre coreferències d'especial interès, ja que sense analitzar-ho profundament, ja existeix una divergència entre text artificial i humà, l'anàlisi de les coreferències permetrà observar més profundament aquesta propietat.

Temps estimat: 25h

Recursos humans: Lingüista computacional

Recursos materials: Ordinador, Servidor

Anàlisi de distàncies respecte als arbres semàntics: T12

Relacionat a l'anàlisi de coreferències,s'ha de realitzar un anàlisi respecte a la distància entre un objecte i la seva anàfora, especialment si, ja que el text artificial te més pronoms, aquesta distància és menor o major respecte al text humà, ja que seria d'esperar que si hi ha més pronoms en el mateix text, aquests tinguin una distància menor, encara així podria ser que es refereixin a diferents objectes, i la distància sigui semblant, per tant un anàlisi és necessari.

Temps estimat: 25h

Recursos humans: Lingüista computacional

Recursos materials: Ordinador, Servidor

Anàlisi de la llei de la brevetat: T13

Es realitzarà un anàlisi sobre la llei de la brevetat [17], és a dir, la relació negativa entre la freqüència d'una paraula amb la seva mida, ja que l'explicació d'aquesta llei és el fet que la gent utilitza paraules més curtes més freqüentant per conservar energia, però un model artificial gasta la mateixa energia, és possible observar propietats especials sobre el text artificial.

Temps estimat: 24h

Recursos humans: Lingüista computacional

Recursos materials: Ordinador, Servidor

Anàlisi de synsets: T14

Es realitzarà un anàlisi sobre els synsets, per obtenir informació sobre el vocabulari utilitzat, si aquest utilitza paraules més genèriques o menys genèriques, ja que si ha après de text variat, es probable que el model prefereixi llenguatge no descriptiu, ja que aquest serà més probable en general.

Temps estimat: 25h

Recursos humans: Lingüista computacional

Recursos materials: Ordinador, Servidor

9.1.3 Implementació d'API d'extracció d'informació

Prototipat de extractor: T15

A causa de la metodologia àgil, i l'èmfasi en tenir extractors d'informació operatius, cada setmana es crearan diferents prototipus de extractors d'informació i s'avaluaran contra les dades existents d'Open-AI, els millors extractors seran guardats per a la creació del classificador definitiu.

Temps estimat: 160h / 10h setmana

Recursos humans: Administrador del projecte, Programador, Lingüista computacional

Recursos materials: Ordinador, Servidor

Implementació de extractor: T16

Una vegada s'hagin acabat l'anàlisi i el prototipat de extractors, s'implementarà la millor solució trobada, utilitzant aquests elements trobats com a rellevants en l'anàlisi, i el millor mètode per a la extracció de l'informació.

Temps estimat: 25h

Dependències: T15

Recursos humans: Programador

Recursos materials: Ordinador

Implementació d'API: T17

Quan el extractor definitiu estigui operatiu, s'integrarà a una simple API rest per a oferir el servei de classificació de text d'una manera simple, independent de plataforma o llenguatge, permitent així la simple integració del classificador amb qualsevol software.

Temps estimat: 24h

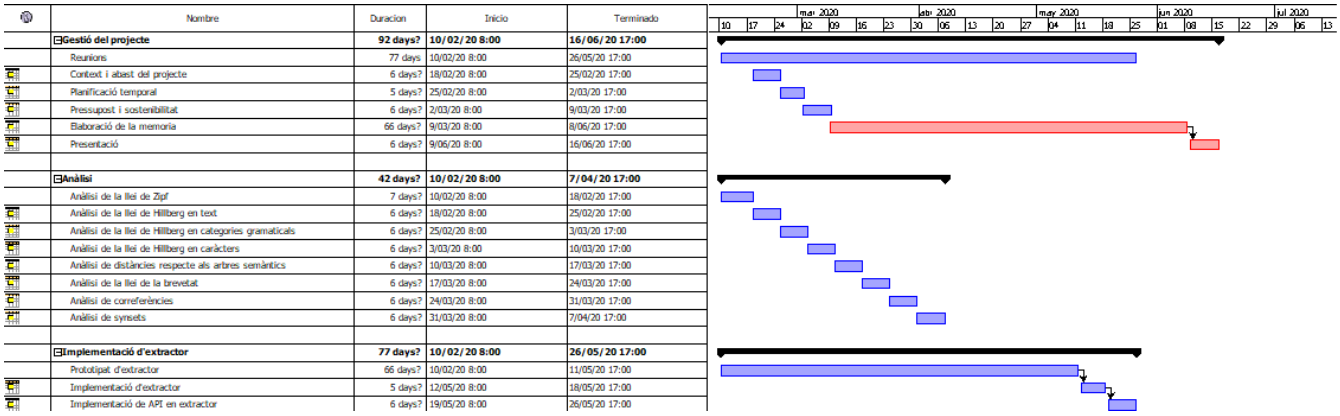
Dependències: T16

Recursos humans: Programador

Recursos materials: Ordinador

9.2 Diagrama de Gantt

Ja que l'estructuració de les tasques segueix una metodologia àgil, el diagrama està dividit per setmanes. Les tasques que es fan en paral·lel al llarg del treball s'han deixat com una sola tasca que s'expandeix per tot el temps que està activa.



	Nombre	Duracion	Inicio	Terminado
	Gestió del projecte	**92 days?**	**10/02/20 8:00**	**16/06/20 17:00**
	Reunions	77 days	10/02/20 8:00	26/05/20 17:00
	Context i abast del projecte	6 days?	18/02/20 8:00	25/02/20 17:00
	Planificació temporal	5 days?	25/02/20 8:00	2/03/20 17:00
	Pressupost i sostenibilitat	6 days?	2/03/20 8:00	9/03/20 17:00
	Elaboració de la memoria	66 days?	9/03/20 8:00	8/06/20 17:00
	Presentació	6 days?	9/06/20 8:00	16/06/20 17:00
	Anàlisi	**42 days?**	**10/02/20 8:00**	**7/04/20 17:00**
	Anàlisi de la llei de Zipf	7 days?	10/02/20 8:00	18/02/20 17:00
	Anàlisi de la llei de Hillberg en text	6 days?	18/02/20 8:00	25/02/20 17:00
	Anàlisi de la llei de Hillberg en categories gramaticals	6 days?	25/02/20 8:00	3/03/20 17:00
	Anàlisi de la llei de Hillberg en caràcters	6 days?	3/03/20 8:00	10/03/20 17:00
	Anàlisi de distàncies respecte als arbres semàntics	6 days?	10/03/20 8:00	17/03/20 17:00
	Anàlisi de la llei de la brevetat	6 days?	17/03/20 8:00	24/03/20 17:00
	Anàlisi de correferències	6 days?	24/03/20 8:00	31/03/20 17:00
	Anàlisi de synsets	6 days?	31/03/20 8:00	7/04/20 17:00
	Implementació d'extractor	**77 days?**	**10/02/20 8:00**	**26/05/20 17:00**
	Prototipat d'extractor	66 days?	10/02/20 8:00	11/05/20 17:00
	Implementació d'extractor	5 days?	12/05/20 8:00	18/05/20 17:00
	Implementació de API en extractor	6 days?	19/05/20 8:00	26/05/20 17:00

Figura 39: Diagrama de Gantt

9.3 Gestió de risc

Similitud entre text automàtic i humà: - Probabilitat mitjana - Impacte alt

Si els anàlisis tornen resultats no suficients, llavors és necessari com a solució agafar els classificadors existents i integrar-los amb el millor classificador obtingut, i d'aquesta fusió mirar quina combinació de classificador basat a votar és la millor. Llavors el classificador encara podria mantenir una mica de generalització a la vegada que obtingués una alta precisió. D'aquesta manera, en el pitjor cas, la precisió només seria lleugerament inferior a l'existent en el discriminador basat en RobERTa [20], però mantindria una major capacitat per discriminar textos no generats per GPT2.

Baix rendiment: - Probabilitat baixa - Impacte baix

Si el classificador té un rendiment molt inferior, llavors primer es buscarà maneres de fer més ràpida l'execució d'aquest, primer buscant ineficiències al codi, i després mirant que és paral·lelitzable d'aquest, si això no és suficient, es buscarien solucions basades en l'arquitectura, intentant desacoblar les parts de càlcul intensives del sistema per a permetre una arquitectura distribuïda, arribant així a la major possible eficiència.

Temps insuficient: - Probabilitat alta - Impacte mitjà

En el cas que cert anàlisi tardi molt temps, primer es projectarà el temps necessari per a l'anàlisi, i si aquesta projecció és excessiva, llavors primer es buscarà fer-lo més eficient, primer buscant ineficiències al codi, i després buscant les parts paral·lelitzables d'aquest. Si encara així no és possible, s'indicarà un temps límit, i el que s'hagi analitzat en aquest temps s'utilitzarà com la solució òptima. Amb l'existència d'un servidor dedicat, és impossible obtenir major capacitat de processament.

Pandèmia: - Probabilitat alta - Impacte baix

Si es declarés pandèmia per l'extensió d'una malaltia, només afectaria les reunions programades, així que s'hauria de modificar com aquestes es porten a cau, utilitzant un sistema que permeti comunicació a distància, com Skype. Aquest risc no posa en perill ninguna altra activitat ja que totes les altres activitats són realitzables en quarantena.

10 Gestió econòmica

10.1 Costos per activitat

El projecte requerirà 3 rols diferents per al seu desenvolupament, aquests són, l'administrador del projecte, que tindrà coneixement en enginyeria del software i es responsabilitzarà de les tasques de disseny i administració, el lingüista computacional, que es responsabilitza de l'anàlisi sobre el text, i el programador, que es responsabilitza de crear els classificadors i integrar aquests en una API.

Per a cada rol necessari per al desenvolupament d'aquest projecte, s'ha estimat una retribució horària utilitzant informació sobre el sou mitja ofert a Espanya, junt amb el cost de la seguretat social que proposaria aquest sou, aquesta estimació és visible a la taula 1.

Rol	**Retribució horària**
Lingüista computacional [2]	18€
Programador [3]	14€
Administrador del projecte [4]	20€

Taula 2: Retribució horària per rol

Utilitzant les estimacions de retribució horària per rol esmentades, el cost de cada tasca és estimat com al cost necessari per al personal que la fa. A la taula 2 es poden veure les tasques dividides en les hores que cada rol ha de fer, i els costos totals associats a cadascuna.

Tasca		**Rol**		**Cost**
	Lingüista computacional	**Programador**	**Administrador del projecte**	
Gestió del projecte	0h	0h	133h	2660€
Reunions	0h	0h	16h	320€
Context i abast del projecte	0h	0h	25 h	500€
Planificació Temporal	0h	0h	8h	160€
Pressupost i sostenibilitat	0h	0h	9h	180€
Elaboració de la memoria	0h	0h	70h	1400€
Presentació	0h	0h	5h	100€
Anàlisi	198h	0h	0h	3564€
Anàlisi de la llei de Zipf	24h	0h	0h	432€
Anàlisi de la llei de Hillberg amb text	25h	0h	0h	450€
Anàlisi de la llei de Hillberg amb categories gramaticals	25h	0h	0h	450€
Anàlisi de la llei de Hillberg amb caràcters	25h	0h	0h	450€
Anàlisi de correferències	25h	0h	0h	450€
Anàlisi de distàncies respecte als arbres semàntics	25h	0h	0h	450€
Anàlisi de la llei de la brevetat	24h	0h	0h	432€
Anàlisi de synsets	25h	0h	0h	450€
Implementació de classificador	10h	187h	12h	3038€
Prototipat de extractor	10h	140h	10h	2340€
Implementació de extractor	0h	25h	0h	350€
Implementació d'API amb extractor	0h	22h	2h	348€
Total	208h	187h	145h	9262€

Taula 3: Costos del personal basats en les tasques de la planificació temporal

10.2 Costos genèrics

10.2.1 Amortitzacions

Tot el software utilitzat pel projecte és d'ús lliure, per tant l'única amortització és en hardware, específicament en el servidor necessari per fer els costosos anàlisis i en l'ordinador necessari per treballar sobre el servidor. El servidor s'assumirà com un simple servidor de torre, específicament un PowerEdge T30. L'ordinador s'estimarà com un ordinador de sobretaula genèric amb un monitor genèric.

Hardware	Preu	Vida útil	Amortització
Servidor	417,45€	5 anys	27,83€
Ordinador	1000€	5 anys	66,7€
Total			94.53€

Taula 4: Cost del hardware

10.2.2 Espai de treball

El cost de l'espai necessari per a portar a terme aquest projecte, seria el d'una oficina compartida simple, que segons [9], costaria uns 200€ al mes, amb la duració esperada del projecte de 4 mesos, això significaria un cost de 800€ per al lloc de treball. El lloc de treball inclouria connexió web.

10.2.3 Consum elèctric

Per calcular els costos elèctrics, aproximem les hores d'ús del servidor, a totes les hores del dia durant tota la duració del projecte, i les hores d'ús de l'ordinador a totes les hores de la duració del projecte. Agafant com el preu de KWh la tarifa oferida per Endesa de 0,119893 €/kWh [10]. Amb una duració estimada de cent tretze dies, i 540 hores de treball resulta en els costos de la taula 4.

Hardware	Potència	Hores	Consum Total	Cost
Servidor	200W	61020h	12204KWh	65,03
Ordinador	200W	540h	108KWh	12,95
Total				77,98€

Taula 5: Cost del consum elèctric

10.2.4 Costos genèrics totals

El cost total dels costos genèrics es correspon a la taula 5.

Activitat	Cost
Amortitzacions	94,53€
Espai de treball	800€
Consum elèctric	77,98€
Total	972,5€

Taula 6: Costos genèrics del projecte

10.3 Costos de contingència

S'estima que la possibilitat que esdeveniments imprevistos apareguin és bastant possible, per això es prepara un fons de contingència del 15%. Amb el cost total brut del projecte, estimat com els costos de personal més els costos genèrics, que són 10234.50€, el fons de contingència seria de 1535.18€.

10.4 Costos per imprevistos

El projecte requereix la capacitat d'aplicar plans alternatius, com va ser observat a la planificació de riscos, per això s'estimaran costos relatius a l'incident possible i el cost de la solució d'aquest incident. El cost serà estimat com a cost del pla alternatiu per la probabilitat que sigui necessari aplicar-lo.

Imprevist	**Cost**	**Probabilitat**	**Cost amortitzat**
Arreclar servidor	200€	5%	10€
Arreclar ordinador	400€	5%	20€
Integrar classificadors existents (20h)	280€	20%	56€
Anàlisis molt costosos en temps (10h)	140€	10%	14€
Total			100€

Taula 7: Cost dels imprevistos

10.5 Costos totals

En la taula 7, s'estima el cost total del projecte, utilitzant les estimacions totals de cada categoria anterior.

Activitat	**Cost**
Cost personal	9262€
Costos genèrics	972,5€
Costos de contingència	1535,18€
Costos per imprevistos	100€
Total	11869,68€

Taula 8: Costos totals del projecte

10.6 Gestió del projecte

Pel control de la gestió del projecte, aprofitant la metodologia àgil seguida, cada setmana es calcularà la desviació entre les hores necessàries per al compliment de la tasca a examinar, contra les hores reals observades, la desviació d'hores seria calculada, per cada tasca com a (hores reals/hores estimades), a la

vegada, es crearia una mètrica complementaria, que seria la desviació de cost, que es calcularia com a ((hores reals-hores estimades)*cost), aquestes dues mètriques s'utilitzen, ja que no tota tasca té el mateix efecte sobre el cost, i no afecta igual més hores de programador, que més hores d'administrador.

Per a la gestió de recursos no relacionats al personal, no s'utilitzarà material no especificat anteriorment si això no és absolutament necessari, és a dir, el projecte no pot progressar sense allò.

Si les desviacions entre valors esperats i valors estimats acaben sent molt elevades, es recalcularà la resta del projecte existent, intentant optimitzar recursos, modificant tasques en progrés i tasques futures per a tornar a crear una estimació amb la qual determinar les desviacions, i que aquesta sigui el més semblant possible a l'antiga planificació.

11 Sostenibilitat i compromís social

11.1 Estudi del impacte ambiental

L'impacte ambiental d'aquest projecte és principalment a les emissions de CO_2 provocades per la generació de l'electricitat necessària per alimentar els ordinadors que són utilitzats per desenvolupar-lo, i els ordinadors on sigui després desplegat per ser utilitzat al llarg de la seva vida útil. Ja que la memòria del projecte serà impresa, existeix el cost en paper d'aquesta, però és un cost negligible.

Respecte al consum del seu desenvolupament, ja que requereix d'anàlisis que requereixen molt temps, necessita l'ús d'un servidor dedicat, aquest servidor estarà operant les vint-i-quatre hores, i, mentre que les CPU modernes tenen sistemes per reduir consum quan la utilització és menor, el consum acumulat segueix sent elevat. El servidor també és utilitzat per altres tasques, així que l'empremta provocada per aquest projecte no és tot el consum del servidor, encara així, per fer una estimació, s'agafa el pitjor cas, on el servidor només és utilitzat per aquest projecte. Assumint 200W com a consumició estàndard del servidor, produiria, al llarg del desenvolupament del projecte, un consum energètic d'uns 4.8KWh al dia, que segons la European Environment Agency [8] que posa les emissions de CO_2 a 265,4g per KWh a Espanya, equival a uns 1273,92g de CO_2 per dia. Al llarg del projecte, això serien cent tretze dies, produint aproximadament uns 144kg de CO_2. També s'ha de considerar el treball fet amb altres ordinadors, ja que sobre el servidor només s'envien les execucions, agafant el mateix consum de 200W, i acotant el temps d'ús a només les hores especificades, tenim un cost energètic de 108KWh i l'emissió aproximada de 29kg de CO_2. En total, el desenvolupament del projecte provocaria l'emissió d'uns 173kg de CO_2.

Ja que els ordinadors utilitzats en el projecte, i el servidor, han estat en ús abans del projecte, i la seva vida útil s'estendrà per sobre de l'extensió del projecte, el consum provocat per la creació del hardware necessari per al projecte es minimitzat.

11.2 Estudi de l'impacte econòmic

Econòmicament, el cost del projecte és relativament elevat, ja que està tractant tecnologia innovadora, on gent amb experiència és escassa, a la vegada, requereix recursos de computació extensius, encara que això s'intenta minimitzar utilitzant hardware més barat, el cost elèctric es manté. Optimitzar aquests és possible, si es redueix el cost del servidor, a llogar un servidor cloud, encara així, els càlculs s'han fet

utilitzant les pitjors possibles projeccions, on el servidor i un ordinador s'han de comprar, i res que no sigui aquest projecte utilitzi el servidor, això dóna una estimació més pessimista de la real, però la real és difícil d'estimar amb precisió, ja que fàcilment es podria estimar per sota de la realitat, així que s'utilitza la visió pessimista com a aproximació.

11.3 Estudi del impacte social

En l'àmbit individual, aquest treball em permetrà explorar en profunditat una gran quantitat de mètodes d'anàlisi de text, mètodes que no són ensenyats al grau d'Enginyeria Informàtica, a la vegada, em permet entendre a major profunditat els models generadors de llenguatge i les seves propietats, això és important, ja que existeix una forta relació entre aquests models i les fake news cosa que em permetrà evitar propaganda amb major facilitat.

Una vegada finalitzat, el projecte oferirà una nova eina per a la extracció de propietats del text artificial, comparats amb les dades utilitzades per altres classificadors existents, aquestes podrien afegir nova informació, que a la seva vegada podria permetre la creació de millors eines, que a la vegada consumeixen menys recursos, permetent una major facilitat per a la integració amb sistemes existents, això ajudaria a lluitar contra fake news generades amb intent de fer propaganda de certes posicions polítiques. Com es put veure a [32], els models generadors de text estan entrant a la cultura popular, i els seus perills comencen a ser visibles al públic, per tant una solució a aquests tindria un impacte social positiu. També a considerar, tota anàlisi fet per aquest projecte, serveix per entendre millor els models generadors de text, cosa que pot ajudar a la millora, o especialització dels models existents, aquests models poden tenir usos positius, com pot ser assistents per escriure, que t'ajuden a escollir la paraula que vols, o fins i tot, traductors, que utilitzen la mateixa tecnologia, però un mètode d'entrenament diferent, però, això suposa que també es podria crear millor text artificial, més difícil de detectar.

12 Conclusions

En aquest treball s'han analitzat diferents lleis i distribucions presents al llenguatge natural i s'han observat les diferencies existents en aquestes comparades amb llenguatge generat per GPT2. A la vegada que s'ha creat una API per obtindre les dades analitzades.

S'ha vist que, mentre que extremadament similar, existeixen diferencies, encara que son microscòpiques, i la seva identificació només es possible quant existeix suficient text tal que la variància es reduïda suficientment tal que la diferencia es suficientment significativa.

Encara així, només amb una petita diferencia, característiques simples poden discriminar amb una precisió de fins a un 71.5%, i la fusió d'aquestes pot arribar a un 74%, quant es tracta amb textos de mida relativament reduïda (com a molt 2048 paraules).

Amb la informació extreta, seria possible enriquir classificadors existents, podent afegir aquesta informació com a metadada del text propi, i a la vegada seria possible crear classificadors més avançats només amb aquestes dades, ja que només amb regressió logística [16] es podia obtenir un 74% de precisió.

També, la creació de la API extractora de dades, permet simplificar i estandarditzar la extracció d'aquestes dades, permetent anàlisis futurs sobre altres models generadors de text, incloent aquests que encara no existeixen

12.1 Treball Futur

Encara que el projecte està acabat, moltes branques futures d'investigació romanen obertes, a més de possibles aplicació d'aquest nou coneixement sobre els textos generats per GPT2.

1. Trobar el punt a partir el qual el text de GPT2 te una variància suficientment reduïda com per a diferenciar-lo de text humà amb facilitat.
2. Generar o enriquir classificadors per diferenciar entre text humà i GPT2 utilitzant els coneixements adquirits en aquest treball.
3. Analitzar més propietats avançades característiques de les distribucions, com el doble règim de la distribució de Zipf.

4. Aplicar aquest anàlisis a altres models generadors de llenguatge, incloent GPT3, utilitzant l'extractor de dades creat durant el projecte.

13 Bibliografia

Referències

[1] URL: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

[2] URL: https://www.linkedin.com/salary/computational-linguist-salaries-in-spain.

[3] URL: https://www.linkedin.com/salary/explorer?countryCode=es&titleId=539.

[4] URL: https://www.linkedin.com/salary/explorer?countryCode=es&titleId=9.

[5] Eduardo G. Altmann i Martin Gerlach. "Statistical Laws in Linguistics". A: *Creativity and Universality in Language* (2016), pàg. 7-26. ISSN: 2195-1942. DOI: 10.1007/978-3-319-24403-7_2. URL: http://dx.doi.org/10.1007/978-3-319-24403-7_2.

[6] *API Development for Everyone*. URL: https://swagger.io/.

[7] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[8] *CO2 emission intensity*. Des. de 2018. URL: https://www.eea.europa.eu/data-and-maps/daviz/co2-emission-intensity-5#tab-googlechartid_chart_11_filters=%7B%22rowFilters%22:%7B%7D;%22columnFilters%22:%7B%22pre_config_ugeo%22:[%22European%20Union%20(current%20composition)%22;%22Spain%22]%7D%7D.

[9] *Coworking*. URL: https://www.spacesworks.com/es/productos-y-servicios/coworking/.

[10] *Endesa One Light*. URL: https://www.endesa.com/ca/cataleg/llum/one/tarifa-one-llum.

[11] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[12] Ramon Ferrer-i-Cancho i Ricard Sol. "Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited". A: *Santa Fe Institute, Working Papers* (gen. de 2000).

[13] Sebastian Gehrmann, Hendrik Strobelt i Alexander M. Rush. "GLTR: Statistical Detection and Visualization of Generated Text". A: *CoRR* abs/1906.04043 (2019). arXiv: 1906.04043. URL: http://arxiv.org/abs/1906.04043.

[14] *GPT-3: A Brief Summary*. Maig de 2020. URL: https://leogao.dev/2020/05/29/GPT-3-A-Brief-Summary/.

[15] Wolfgang Hilberg. "Der bekannte Grenzwert der redundanzfreien Information in Texten - eine Fehlinterpretation der Shannonschen Experimente?" A: *Frequenz* 44 (set. de 1990), pàg. 243 - 248. DOI: 10.1515/FREQ.1990.44.9-10.243.

[16] David W. Hosmer i Stanley Lemeshow. *Applied logistic regression*. John Wiley i Sons, 2000. ISBN: 0471356328, 9780471356325.

[17] Clyde Kluckhohn. "Human Behavior and the Principle of Least Effort. George Kingsley Zipf". A: *American Anthropologist* 52.2 (1950), pàg. 268 - 270. DOI: 10.1525/aa.1950.52.2.02a00290. eprint: https://anthrosource.onlinelibrary.wiley.com/doi/pdf/10.1525/aa.1950.52.2.02a00290. URL: https://anthrosource.onlinelibrary.wiley.com/doi/abs/10.1525/aa.1950.52.2.02a00290.

[18] J. Lin. "Divergence measures based on the Shannon entropy". A: *IEEE Transactions on Information Theory* 37.1 (1991), pàg. 145 - 151.

[19] Haitao Liu, Chunshan Xu i Junying Liang. "Dependency distance: A new perspective on syntactic patterns in natural languages". A: *Physics of Life Reviews* 21 (març de 2017). DOI: 10.1016/j.plrev.2017.03.002.

[20] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". A: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

[21] Michael Mauldin. "Semantic Rule Based Text Generation". A: (maig de 2002). DOI: 10.3115/980431.980568.

[22] *Natural Language Toolkit¶*. URL: https://www.nltk.org/.

[23] *News You Can't Use*. URL: https://newsyoucantuse.com/.

[24] Openai. *openai/gpt-2-output-dataset*. Des. de 2019. URL: https://github.com/openai/gpt-2-output-dataset.

[25] Openai. *openai/gpt-2-output-dataset*. URL: https://github.com/openai/gpt-2-output-dataset/blob/master/detection.md.

[26] Alec Radford. *Better Language Models and Their Implications*. Des. de 2019. URL: https://openai.com/blog/better-language-models/.

[27] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". A: 2019.

[28] Irene Solaiman. *GPT-2: 1.5B Release*. Nov. de 2019. URL: https://openai.com/blog/gpt-2-1-5b-release/.

[29] *spaCy · Industrial-strength Natural Language Processing in Python*. URL: https://spacy.io/.

[30] *Transformer-XL: Unleashing the Potential of Attention Models*. Gen. de 2019. URL: https://ai.googleblog.com/2019/01/transformer-xl-unleashing-potential-of.html.

[31] Ashish Vaswani et al. "Attention Is All You Need". A: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[32] Vox. *How robot writers could change the internet*. Març de 2020. URL: https://www.youtube.com/watch?v=gcHkxP9adiM.

[33] Thomas Wolf. *State-of-the-art neural coreference resolution for chatbots*. Oct. de 2017. URL: https://medium.com/huggingface/state-of-the-art-neural-coreference-resolution-for-chatbots-3302365dcf30.

[34] Viatcheslav Yatsko. "Automatic text classification method based on Zipf's law". A: *Automatic Documentation and Mathematical Linguistics* 49 (juny de 2015), pàg. 83 - 88. DOI: 10.3103/S0005105515030048.

[35] Shuiyuan Yu, Chunshan Xu i Haitao Liu. "Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation". A: *CoRR* abs/1807.01855 (2018). arXiv: 1807.01855. URL: http://arxiv.org/abs/1807.01855.

[36] George K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, 1949.